Métodos estadísticos para tratar incertidumbre en estudios de asociación genética: aplicación a CNVs y SNPs imputados

Isaac Subirana Cachinero

Métodos estadísticos para tratar incertidumbre en estudios de asociación genética: aplicación a CNVs y SNPs imputados

Isaac Subirana Cachinero.
Barcelona, julio de 2014.

Métodos estadísticos para tratar incertidumbre en estudios de asociación genética: aplicación a CNVs y SNPs imputados

Memoria presentada por **Isaac Subirana Cachinero** para optar al grado de Doctor por la Universitat de Barcelona

Programa: Estadística.
Departamento: Estadística, Facultad de Biología.
Centro de realización de la tesis: IMIM - Parc de Salut Mar.
Fecha: julio de 2014, Barcelona.

Isaac Subirana (Doctorando)
CIBER en Epidemiología y Salud Pública (CIBERESP)
IMIM - Parc de Salut Mar
Departamento de Estadística, Universitat de Barcelona

Dr. Juan Ramón González (Director)
Centro de Investigación en Epidemiología Ambiental (CREAL)
CIBER en Epidemiologia y Salud Pública (CIBERESP)
Departamento de Matemáticas, Universitat Autònoma de Barcelona

Dr. Josep Maria Oller (Tutor)
Departamento de Estadística, Universitat de Barcelona

Dr. Antoni Monleón Getino (Co-director)
Departamento de Estadística, Universitat de Barcelona

Impresa en junio de 2014
con el apoyo de la Fundación IMIM

Agradecimientos

Quiero agradecer a Juan Ramón González por su ímpetu y energía en la dirección de esta tesis, desde el principio y a lo largo de todos los trabajos que la conforman. Haber trabajado con él todo este tiempo ha sido sin duda una experiencia muy productiva y fructífera para mi carrera como técnico investigador, más allá del ámbito estricto de esta tesis doctoral. A Mikel Esnaola, por su paciencia e inmensa ayuda que me ofreció con una parte muy importante de la tesis y con quien he aprendido muchísimo. También quiero agradecer a los compañeros de la URLEC, especialmente a Joan Vila por su inestimable apoyo durante la elaboración de la tesis así como sus comentarios en la memoria que han contribuido mucho en su mejora. A mi codirector, Antoni Monleón. A mi familia, y en general a todas las personas que me han apoyado durante todos estos años.

Índice general

INTRODUCCIÓN

1.1. Resumen

En los últimos años, se han descubierto un gran número de variantes genéticas de distinta naturaleza, desde las más simples que indican un cambio en un nucleótido, *Single Nucleotid Polymorphism* (SNP), hasta otras más complejas referentes al número de copias de un segmento de la cadena de ADN, *Copy Number Variant* (CNV). Existen otras variantes como son las inversiones, microsatélites, etc. Sin embargo, esta tesis se ha focalizado en los SNPs y en los CNVs, ya que son los dos tipos de variantes más analizadas en los estudios de epidemiología genética.

En muchas situaciones, los métodos para analizar el efecto que tienen los SNPs o los CNVs sobre las enfermedades están bien resueltos. Sin embargo, en algunos casos, los SNPs y los CNVs se observan con incertidumbre. Por ejemplo, a veces el genotipo para un SNP no se observa directamente sino que se imputa. A su vez, establecer el número de copias para un CNV se hace de forma indirecta a partir de la señal cuantitativa de su sonda (*probe*). Esto hace que se requieran métodos estadísticos "no estándar" apropiados para estudiar la asociación entre SNPs imputados o CNVs incorporando esta incertidumbre.

En la literatura se han descrito diferentes estrategias para afrontar los estudios de asociación entre una variante genética medida con incertidumbre y una variable respuesta: (i) la estrategia Naive y (ii) la estrategia conocida como Dosage. A *grosso modo*, la primera no tiene en cuenta la incertidumbre, mientras que la segunda lo hace de forma aproximada.

En esta tesis doctoral se presentan métodos estadísticos para tratar datos medidos con incertidumbre que solventen las limitaciones que presentan los métodos existentes. Esta aportación es relevante y lo será más en el futuro ya que cada vez se analizan variantes genéticas medidas con incertidumbre (CNVs, SNPs imputados, variantes obtenidas a partir de datos de ultrasecuenciación, inversiones, ...). Los estudios genéticos suelen analizar el genoma completo (estudios conocidos como GWAS, por sus siglas en inglés - Genome Wide Association Study) y miles de participantes. Mientras el *software* y/o el *hardware* no aporten nuevas soluciones es poco eficiente utilizar una técnica precisa y potente que es poco eficiente computacionalmente para dar solución a este tipo de estudios. La solución es una técnica de análisis de variantes genéticas que incorporen esta incertidumbre.

Se han propuesto y descrito analíticamente modelos estadísticos para estudiar la asociación entre variantes genéticas medidas con incertidumbre y una variable respuesta. Dichos modelos tienen la característica de incorporar la incertidumbre de forma adecuada en la función de verosimilitud. Se ha considerado los siguientes escenarios, entre otros: que la variable respuesta sea de tipo binario (presencia o no de cierta enfermedad), cuantitativa (nivel de colesterol en sangre) ó censurada (tiempo hasta recaída). No sólo se han diseñado técnicas para el análisis de las variantes genéticas de forma individual sino también para pares simultáneamente (interacciones). Todo ello se ha implementado en distintas funciones estructuradas e integradas como parte de un programa de uso común en la epidemiología genética como es `R`. Además se ha escrito parte del código de las funciones en lenguaje `C++` a fin de que los cálculos sean mucho más rápidos.

1.2. Estudios de asociación genética

Gracias a la mejora tecnológica, durante los últimos años se han descubierto un gran número de estructuras genéticas [1], lo cual ha sido sido de vital importancia para entender mejor las diferencias entre individuos y predecir cuáles son más susceptibles de padecer ciertas enfermedades [2, 3, 4, 5].

En los primeros años, los estudios de enfermedades o rasgos hereditarios, se llevaban a cabo en familias [6, 7], y permitían saber qué grado de la enfermedad era explicado por la genética. No obstante, su limitación radicaba en la dificultad de precisar las variantes genéticas y su estructura implicadas en la enfermedad. Además, aunque tuvieron éxito en identificar rasgos con gran carga genética (o penetrancia) no fue así para rasgos en que la genética explicaba una proporción más limitada como suele ser el caso de enfermedades complejas (por ejemplo gran variedad de tipos de cáncer o patologías cardiovasculares). A pesar de que los estudios en familias se siguen llevando a cabo [8], éstos dieron paso a los estudios de asociación donde los individuos participantes no están emparentados y en que típicamente la muestra es mucho mayor. Naturalmente, los estudio de asociación son posibles gracias a la vertiginosa disminución de los costes en el proceso de genotipado [9].

Con el incremento del tamaño de la muestra se ganó en potencia estadística y fue más fácil identificar variantes genéticas aunque su asociación con la enfermedad fuera más pequeña (baja penetrancia). Además, los estudios de asociación fueron capaces de recopilar centenares de miles de variantes genéticas, donde se estudiaban desde cambios de un sólo nucleótido (SNPs) [10, 11, 12] a eliminaciones o repeticiones de segmentos de varios miles de nucleótidos (CNVs) [13, 10, 14], y otro tipo de variantes como las inversiones de segmentos [15], etc.

El gran número de variantes genéticas identificadas, lejos de ser una ventaja des del punto de vista estadístico "tradicional", supone un reto para los estudios de asociación. Para su análisis existen diferentes estrategias. La más simple es estimar la asociación de cada variante una a una por separado con la variable respuesta,

lo cual se traduce en ajustar el mismo modelo de asociación repetidamente miles o centenares de miles de veces. Esta estrategia es la que se suele realizar en la gran mayoría de estudios de asociación donde se analizan hasta millones de variantes genéticas a lo largo de todo el genoma (*Genome Wide Association Study* (GWAS)). Otro tipo de estrategia más compleja consiste en involucrar grupos de variantes a la vez. El ejemplo más sencillo de ello es el estudio de interacciones de pares de variantes conocido también como epistatis [16]. En esta tesis nos hemos limitado a los modelos de asociación entre las variantes una a una y la enfermedad y los estudios de interacciones de pares de variantes de SNPs.

1.2.1. *Single Nucleotide Polymorphisms* (SNPs)

Los SNPs fueron las primeras variantes genéticas analizadas en los estudios de asociación, debido a que son las más simples. En la literatura hay un gran número de artículos publicados reportando SNPs asociados con distintas enfermedades, como son el Alzheimer [10], el cáncer [11], o las enfermedades cardiovasculares [12], entre otras

Un SNP se define como la mutación o cambio producido en una determinada posición ó nucleótido ("locus") de la cadena de ADN que ha sido favorecida en términos evolutivos por lo que lo presenta una proporción "significativa" de individuos de la población. Normalmente, se toma una proporción superior al 1 % para que sea considerada como SNP, y si es inferior se define como mutación. Aunque estos límites son arbitrarios. A menudo, cuando la proporción es inferior a 1 % simplemente se llaman *rare* SNPs, y *common* SNPs en caso contrario.

Debido que los seres humanos, como la mayoría de seres vivos, tenemos dos cadenas de ADN (cada una proveniente de uno de los dos progenitores), puede haber un cambio en un *locus* (alelo) en las dos cadenas, en una de ellas o en ninguna. Por lo tanto, un SNP es una variable discreta de 3 posibles valores (0, 1 ó 2). El efecto que puede tener un SNP sobre un rasgo (modelo de herencia) puede ser de varios tipos.

Los más comunes son: dominante, recesivo, aditivo o co-dominante. En el primer caso, los individuos se pueden agrupar entre aquellos que tienen algún cambio ('1' ó '2') frente aquellos que no tienen ninguno ('0'). En el segundo caso, recesivo, los individuos se agrupan entre aquellos en que el cambio se ha producido en las dos cadenas ('2') y los demás. Para el modelo aditivo, en cambio, se entiende que hay un incremento progresivo en el riesgo ó en los valores medios de la variable respuesta al incrementar el número de alelos. Por último, el modelo codominante, no hace ninguna suposición y trata la variable SNP como una variable categórica de 3 niveles.

1.2.2. *Copy Number variants* (CNVs)

Los CNVs son segmentos largos (típicamente de miles de nucleótidos de longitud) de ADN que se repiten en distintas zonas de la cadena, contiguas o no. Teóricamente, sus posibles valores (o genotipos) son los valores enteros desde cero a infinito. No obstante, en la práctica se suelen distinguir tres situaciones: (i) deleción: cuando un segmento de una determinada zona del genoma que está en la mayoría de los individuos de la población no aparece; (ii) *wild type*: cuando el segmente está presente; (iii) duplicación: cuando el segmento está repetido. Así, el CNV se puede tratar de forma similar a un SNP (con 0, 1 ó 2 copias), y los modelos de herencia pueden ser igualmente aditivo, recesivo, dominante o codominante.

En [17], McCarrol manifiesta la importancia de estudiar estas variantes, más allá de los SNPs, para poder entender las bases genéticas de enfermedades con cierta componente hereditaria. Además, algunos estudios recientes han descubierto CNVs asociados a ciertas enfermedades [13, 10, 14].

1.2.3. *Genome Wide Association Studies* (GWAS)

Los GWAS son estudios de asociación aplicados a la población cuyo objetivo es averiguar qué variantes genéticas a lo largo de **todo** el genoma están asociados a una determinada enfermedad.

Gracias al proyecto HapMap (www.hapmap.org) [18, 19] se han descubierto hasta la fecha más de cuatro mil millones de SNPs. Una de las ventajas que ha aportado el proyecto HapMAP a nivel práctico es que ha evidenciado que no es necesario genotipar todos los SNPs ya que muchos de ellos están en desequilibrio de ligamiento, *Linkage Desequilibrium* (LD), el cual representa la correlación entre un par de SNPs. Se ha visto que genotipando una pequeña proporción de ellos, *tag SNPs* considerados como "representantes" de una zona del genoma donde los SNPs están en alta correlación, es suficiente [20]. Así, se han diseñado plataformas, como Illumina © [21] ó Affymetrix © [22, 23] para genotipar los *tag SNPs* estratégicamente situados cubriendo más del 80 % de la variabilidad explicada por los cuatro mil millones de SNPs. Actualmente estas plataformas son capaces de genotipar más de un millón de SNPs.

En los GWAS, el número de variantes genéticas estudiadas fácilmente sobrepasa el millón cuando se trata de SNPs ó miles para los CNVs. Además, no se hace ninguna suposición *a priori* sobre qué variantes genéticas son más susceptibles de estar asociadas con la enfermedad, a diferencia de los estudios de genes candidatos donde se estudian, como máximo decenas de SNPs. Por este motivo es necesario situar el nivel de significación estadística en valores mucho más pequeños, para mantener el error de tipo I. Es decir, corregir por comparaciones múltiples. En contrapartida, para conseguir una potencia estadística suficiente se requiere un tamaño de muestra mucho mayor. En los estudios actuales se reclutan decenas de miles de participantes. A menudo, para conseguir suficiente tamaño de muestra, se constituyen grandes consorcios para unir diferentes estudios, y en que ellos analizan los datos de cada centro por separado para posteriormente meta-analizar los resultados [24, 25].

1.2.3.1. Estudios caso-control

Los estudios de tipo caso-control son los más comunes en los GWAS ya que son relativamente fáciles de llevar a cabo des del punto de vista logístico a la hora de conformar la muestra. El objetivo de este tipo de estudios es comparar la distribución de las variantes genéticas entre los casos (individuos que presentan la enfermedad) y los controles (individuos sanos o que no presentan la enfermedad). Su diseño tiene la

ventaja, a diferencia de los estudios de cohorte, de que se puede conseguir un gran número de casos con facilidad. Mientras que los controles se seleccionan a partir de una población libre de la enfermedad comparable en edad y sexo ó alguna otra característica que se quiera tener en cuenta (estrategia *matching*).

Las herramientas y modelos estadísticos usados para estimar la asociación entre las variantes genéticas y la enfermedad en este tipo de estudios suelen ser muy simples: cuando no se tiene en cuenta ninguna variable confusora o de ajuste, simplemente se calcula el estadístico χ^2 ó el estadístico exacto F de Fisher [26, 27]; y cuando se tienen en cuenta posibles variables confusoras, como por ejemplo la raza ó las componentes genéticas para controlar por el *population stratification* [28], se suele aplicar una regresión logística donde la variable respuesta es la presencia/ausencia de enfermedad y la asociación con las/s variante/s genética/s es ajustada por variables potencialmente confusoras de esta relación [12].

Desde el punto de vista epidemiológico, los estudios caso-control aportan un grado de evidencia "bajo" en cuanto a la relación causal entre la variante genética y la enfermedad debido a que son estudios retrospectivos o transversales. No obstante, son de gran utilidad para descubrir potenciales variantes genéticas asociadas de forma causal. Para posteriormente asegurarse o tener más evidencia de que la variante genética está realmente asociada de forma causal es necesario realizar un estudio de seguimiento prospectivo (estudios de cohorte).

1.2.3.2. Estudios de cohorte

Los estudios de cohorte, a diferencia de los de caso-control, suelen ser más costosos ya que requieren de un seguimiento a lo largo del tiempo para observar qué individuos padecen la enfermedad en el periodo de estudio. Además de identificar los individuos que padecen la enfermedad, también se registra en qué momento aparece ésta. En estos estudios casi siempre la variable respuesta es "tiempo hasta el evento", si desarrolla el evento o "tiempo hasta la censura" si acaba el seguimiento y no ha desarrollado el evento.

De entre todos los tipos de estudio observacionales, o sea, exceptuando los ensayos clínicos, los estudios de cohorte son los que aportan más evidencia "causal" a la posible asociación entre las variantes genéticas y la enfermedad. El motivo es que, a diferencia de los estudios caso-control, los casos no son elegidos de forma artificial (por diseño) sino que forman parte de la cohorte inicial que idealmente es representativa de la población general. Por otra parte, la variable respuesta aporta más información que en los estudios caso-control ya que no se trata de una variable binaria (caso ó control) sino que es realmente una variable continua, y por lo tanto tienen más potencia estadística para detectar asociación.

Así pues, no es de extrañar que en muchas ocasiones se diseñe y analice un estudio caso-control para "descubrir" posibles variantes genéticas asociadas a una determinada enfermedad y posteriormente estimar y/o ratificar esta asociación en un estudio de cohorte.

Finalmente, el modelo estadístico más usado para analizar un estudio de cohorte con la variable respuesta del tipo tiempo hasta evento con posible censura es la regresión semiparamétrica de Cox de riesgos proporcionales o la regresión paramétrica de Weibull.

1.2.3.3. Estudios de respuesta continua

Los estudios genéticos de población conocidos como estudios de 'respuesta continua' (ó rasgo continuo) son aquellos en que la enfermedad o variable respuesta es de carácter cuantitativo y que además se distribuye según una ley normal, previa transformación si es necesarios (p.e. logaritmo). Un ejemplo de este tipo de estudio se da cuando se estudia los niveles de colesterol, el índice de masa corporal o la tensión arterial, por ejemplo.

Lo estudios en los que la respuesta es del tipo "continua" son transversales, es decir que los participantes son una muestra representativa de la población de estudio a los que se le mide la variable respuesta en un determinado momento del tiempo (el

mismo para todos), a diferencia de los estudios de cohorte que requieren un tiempo de seguimiento, o los estudios caso-control donde los casos y los controles pueden venir de dos fuentes o muestras distintas posiblemente reclutadas en diferentes periodos.

Al tratarse de una respuesta continua, estos estudios tienen una elevada potencia estadística, es decir, es más probable detectar asociación para aquellas variantes genéticas verdaderamente asociadas con la variable respuesta.

Una ventaja compartida con los estudios de cohorte, es que se puede analizar la asociación con distintas variables respuesta, siempre y cuando en el seguimiento se hayan recopilado diferentes tipos de eventos. Por ejemplo, el índice de masa corporal o el colesterol total en los estudios transversales, o ictus o enfermedad coronaria en los estudios de cohortes.

1.3. Incertidumbre en las variantes genéticas

La incertidumbre ocurre cuando en alguno o en todos los individuos de la muestra no se ha podido observar directamente el valor de la variante genética, sino que se tiene una información parcial sobre qué valor o valores ha tomado. Dependiendo del tipo de variante y de los métodos técnicos para medirla (genotipado), la información que se tiene es más o menos difusa. Estadísticamente hablando, este hecho puede ser crítico en el uso de herramientas para describir su distribución en la muestra o hacer inferencia sobre la población, o en último término estimar su asociación con la enfermedad. Por ello, es necesario evaluar la fiabilidad de las herramientas existentes para llevar a cabo estos objetivos y, si es necesario, considerar herramientas o modelos alternativos.

1.3.1. SNPs imputados

En los últimos años, y con la mejora de la tecnología, los estudios genéticos de población han incorporado un número creciente de SNPs, eso es, se han genotipado más variantes genéticas de este tipo en una misma muestra, creciendo su número de forma exponencial: se ha pasado de genotipar desde pocas decenas de SNPs hasta un millón.

Por otro lado, y gracias al proyecto HapMap , se conocen mejor los patrones que relacionan los diferentes SNPs entre ellos. Y se sabe qué regiones o conjuntos de SNPs están más relacionados (bloques en LD). Esto es de gran utilidad, primero porque no hace falta genotipar todos los SNPs, y segundo porque es posible "inferir" SNPs que no se hayan genotipado pero que están relacionados con otros a su alrededor que sí se hayan observado. Usando los patrones de LD del HapMap, se pueden inferir, o mejor dicho, imputar hasta 1,5 millones de SNPs a partir de 1 millón de genotipados. Esto hace posible incorporar muchas más variantes genéticas en un GWAS a un coste adicional cero. Otra ventaja es que permite analizar muestras genotipadas con distintas plataformas (Affymetrix © o Illumina © , por ejemplo) en que se han observado conjuntos de SNPs diferentes, mediante un meta-análisis. Gracias a la

imputación, el número de SNPs que se pueden estudiar es mucho mayor que el que se tendría si sólo se tienen en cuenta los SNPs genotipados en todas las plataformas (intersección).

Dependiendo del SNP a imputar, sobretodo del grado de LD con otros SNPs genotipados, la incertidumbre será mayor o menor. Aunque hay distintos algoritmos para realizar la imputación (MACH [29], IMPUTE [30], PLINK [31],...) y que en general pueden dar resultados numéricamente distintos, [32], todos ellos reportan las probabilidades de tener 0, 1 ó 2 copias para cada variante imputada y para cada individuo de la muestra. Visto así, los valores de los SNPs imputados ya no son 0, 1 ó 2, sino un vector de 3 probabilidades. Se entiende que hay menos incertidumbre cuando estas probabilidades están cercanas a 0 ó 1. Y sabiendo que la suma de las 3 debe ser 1, se puede decir que habrá poca incertidumbre si la probabilidad máxima de las 3 es próxima a 1 y gran incertidumbre si cada una de las tres probabilidades es próxima a 1/3. Finalmente, se puede tomar el promedio de este índice para todos los individuos. No obstante, esta medida no es la que calculan la mayoría de programas de imputación. Por ejemplo, MACH ó IMPUTE reportan un índice llamado R^2 cuyo rango va de 0 a 1, siendo 0 el grado máximo de incertidumbre y 1 la máxima certeza en la imputación.

La ventaja de asignar un valor numérico a cada SNP imputado permite ordenarlos según el grado de incertidumbre y si se cree necesario eliminar aquellos SNPs con un grado de incertidumbre demasiado alto a partir de un determinado punto de corte. Sin embargo, no está claro cuál debe ser este punto de corte, y ésto es crítico porque existe el peligro de descartar demasiados SNPs del análisis por un lado o de estudiar SNPs con demasiada incertidumbre para que los resultados sean fiables por el otro.

En [33] se describen y comparan con detalle los distintos métodos de imputación y los distintos índices para medir el grado de incertidumbre, así como las estrategias para analizar distintas plataformas en la combinación de diferentes cohortes (meta-análisis).

1.3.2. CNVs

En los CNVs, la incertidumbre viene dada directamente por la técnica de determinación del número de copias. Los CNVs no se observan directamente sino que se infieren a partir de una cantidad numérica observada de la intensidad de unas sondas fluorescentes diseñadas para adherirse al segmento a estudiar. La idea es que cuantas más sondas se adhieran, más intensa es la señal luminosa que se desprende y probablemente más copias del segmento tenga el individuo. Debido a que los segmentos están constituidos por miles de bases (nucleótidos), puede haber sondas que no se adhieran correctamente y otras que directamente no lo hagan. Esto conlleva un cierto grado de error, o sea, incertidumbre.

Después de un proceso de normalización de la intensidad de las sondas, se estudia la distribución de todos los individuos de la muestra y se intenta clasificarlos en distintos grupos según el número de copias. Bajo el punto de vista estadístico, se trata de un problema de *clustering* a partir de una variable cuantitativa. Si la señal sigue una distribución normal, se pueden aplicar técnicas de inferencia basadas en mixturas de normales usando el algoritmo *Expectation-Maximization* (EM) [34]. Alternativamente, si la distribución no sigue una ley normal, se utilizan técnicas más robustas par este tipo de distribuciones [35].

Al igual que para los SNPs imputados, existen distintas medidas de incertidumbre dependiendo del algoritmo usado para inferir el número de copias. Asimismo, también se suelen determinar puntos de corte eliminando los CNVs con demasiada incertidumbre como proceso de control de calidad [12].

1.3.3. Estudios de asociación con incertidumbre

Es importante tener en cuenta la posible incertidumbre a la hora de estimar la asociación entre la variante genética y la variable respuesta, sea cual sea el tipo de estudio (caso-control, cohorte o de respuesta cuantitativa) o el tipo de variante (SNPs imputados o CNVs). No hacerlo puede llevar a resultados sesgados y a tener poca potencia estadística para detectar diferencias significativas [36, 37]. Tampoco

está muy claro cómo incorporar esta incertidumbre en los estudios de asociación, como se apunta en [38]. Por lo tanto, es de suma importancia saber qué grado de error y bajo qué escenarios usar una técnica inapropiada puede dar resultados demasiado erróneos, y cuándo las técnicas estadísticas que no tienen en cuenta la incertidumbre están demasiado sesgadas o por el contrario cuándo son suficientemente válidas.

1.3.3.1. Estrategias generales

Las técnicas estadísticas "clásicas" de asociación de la forma "respuesta $\sim$ variables explicativas" suponen que todas las variables, tanto la respuesta como la/s variables explicativas en las que se incluyen lógicamente las variantes genéticas (SNPs ó CNVs) están observadas para todos los individuos que conforman la muestra. En el caso que no se sepa el valor de alguna variable pero se sepa su distribución, se pueden incorporar las probabilidades en el modelo a partir del teorema de probabilidades totales: la verosimilitud para un individuo será la suma ponderada sobre todos los posibles valores que pueda tomar la variable:

$$L(Y|\Theta) = \sum_k L(Y|X = k; \Theta)P(X = k) \tag{1.1}$$

donde Y es el valor de la variable respuesta, X es la variante genética no observada, Θ es el vector de parámetros del modelo y $P(X = k)$ es la probabilidad de que la variante genética tome el valor k. El vector de probabilidades $P(X = k)$ se obtiene a partir de los algoritmos de imputación en el caso de los SNPs o de la distribución de la intensidad para los CNVs.

Se puede observar que cuando la incertidumbre es nula, la ecuación 1.1 se simplifica a un sólo sumando. Así, la función de verosimilitud es mucho más simple y de hecho el modelo equivale a un modelo de regresión "clásica" (regresión lineal, logística, de Cox ó de Weibull según sea la distribución de la variable respuesta). Pero cuando hay incertidumbre, el sumatorio en 1.1 no se puede simplificar y la estimación de los parámetros por máxima verosimilitud es más compleja. Consecuentemente, es necesario aplicar técnicas de optimización numérica incluso cuando la respuesta se

ajusta a una distribución 'normal'.

Por otro lado, se podría tener en cuenta más de una variante con incertidumbre a la vez. La misma fórmula 1.1 es válida si se interpreta X como la variable resultante de la combinación de todos los posibles valores de las variantes y k sus combinaciones posibles. Esto hace que el número de sumandos en 1.1 se incremente exponencialmente y la complejidad del modelo sea demasiado elevada. Por este motivo, esta tesis se ha restringido a los modelos de asociación genética de una variante o de dos variantes con una interacción como máximo. A la práctica, pero, no es una gran limitación ya que estos son, con diferencia, los estudios más frecuentemente realizados.

Cabe notar, también, que la estrategia descrita en este apartado es válida siempre y cuando las variables explicativas con incertidumbre sean cualitativas, dicotómicas o discretas con un número finito de posibles valores. Es fácil darse cuenta, pues, que la misma argumentación hecha para SNPs o CNVs se podría aplicar también a otras variantes genéticas o incluso a variables no genéticas.

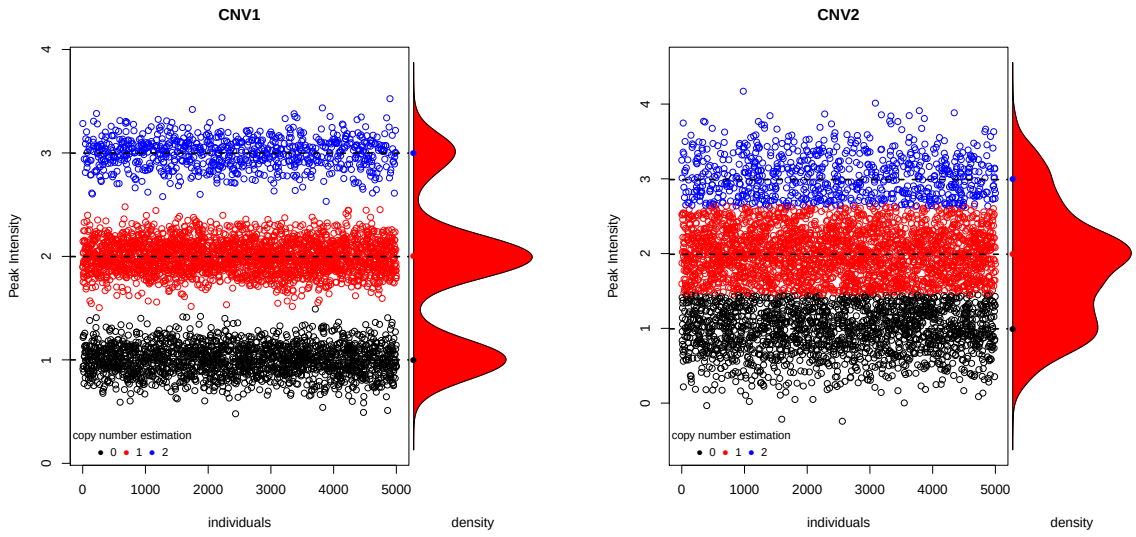
Finalmente, es importante recalcar que esta tesis se ha centrado en los modelos de asociación y no en los algoritmos de imputación de SNPs o de inferencia del número de copias de los CNVs. El modelo estadístico de asociación propuesto tiene el reto de incorporar de forma apropiada la incertidumbre, eso sí, habiendo hallado/estimado previamente el vector de probabilidades $P(X = k)$ para cada individuo de la muestra. Aunque existen otras estrategias en que se estiman estas probabilidades y se ajusta el modelo de asociación en un solo paso [39, 40], éstas no se han considerado en esta tesis. El motivo es que estos algoritmos son mucho más complejos siendo infactibles cuando se estudian muchas variantes genéticas (por ejemplo en un GWAS).

1.3.3.2. Aplicación a los CNVs

Como se ha mencionado anteriormente, el número de copias de un CNV se infiere a partir de una medida continua como es la intensidad de la señal de las sondas. Existen diferentes técnicas para ello (poner puntos de corte "a ojo" a partir de la

distribución o simplemente fijar puntos arbitrarios "a priori") a fin de separar los distintos grupos de la muestra según el número de copias. Otra estrategia existente más sofisticada consiste en usar técnicas estadísticas de discriminación (clúster jerárquico o no jerárquico, o la mixtura de normales, etc.). Aunque esta última estrategia tiene un criterio estadístico, no deja de asignar a cada individuo el número de copias (su valor) sin tener en cuenta la posible incertidumbre en esta asignación.

En la siguiente figura se ilustran las intensidades de dos CNVs distintos con 0, 1 y 2 copias posibles, donde se puede apreciar el diferente grado de incertidumbre. Claramente se ve como el CNV2 tiene mucha más incertidumbre ya que las intensidades de los diferentes grupos están más solapadas y es más difícil clasificar a los individuos, aunque, como se puede apreciar, los puntos de corte resultantes (tanto si se hace "a ojo" como si se usan técnicas estadísticas discriminantes) son los mismos.



Con esta incertidumbre es necesario no sólo asignar el número de copias, sino tener presente esta incertidumbre. Esto se consigue en considerar las probabilidades obtenidas en la inferencia del número de de copias y utilizando una herramienta estadística apropiada que sea capaz de calcular la probabilidad de tener un determinado número de copias para cada individuo dada la intensidad de la señal de la sonda. Por último, y una vez obtenidas estas probabilidades, se puede construir la función de verosimilitud como 1.1, donde X es el CNV estudiado.

1.3.3.3. Aplicación a los SNPs imputados

Para los SNPs imputados, la estrategia es la misma que para los CNVs. Sólo que ahora las probabilidades $P(X = k)$ se obtienen a partir de algún algoritmo de imputación como MACH, IMPUTE ó PLINK que proveen estas probabilidades y a continuación se ajusta la función de verosimilitud donde X es el SNP imputado a estudiar. En el proceso de imputación, si el SNP a imputar está en una zona con poco LD a los SNPs de su alrededor no estará claro qué genotipo (0, 1 ó 2 copias del alelo de riesgo) tendrá el individuo y por lo tanto las probabilidades serán relativamente bajas (próximas a 1/3).

El siguiente paso será ajustar un modelo de asociación entre el SNP y la variable respuesta que incorpore de alguna manera esta incertidumbre, de la misma forma que se hace con los CNVs.

1.4. Creación de paquetes en R

En la actualidad, existen diferentes *softwares*, entre ellos PLINK [31], para llevar a cabo un análisis de GWAS con SNPs o con CNVs. Muchos o la mayoría de ellos están escritos en C++ y no están integrados en ningún programa demasiado "amigable" con el usuario, sino que suelen ser bastante rígidos en el sentido que no permiten manipular los datos de "entrada" y éstos deben estar en un formato específico para que sean leídos. Además, los resultados son más bien difíciles de leer y no son "personalizables".

Como alternativa, en esta tesis, se ha considerado implementar el modelo estadístico para el estudio de asociación genética de variantes con incertidumbre, basado en la verosimilitud presentado en el sección anterior, dentro del *software* estadístico R [41]. Las ventajas de incorporar la herramienta en R son varias, entre ellas: (i) en la actualidad hay muchos usuarios de este programa, (ii) es muy flexible y potente a la hora de leer datos y obtener resultados a "medida", (iii) un usuario familiarizado con R fácilmente puede modificar algunas opciones por su cuenta ya que el código es abierto, (iv) es posible incorporar funciones para representar gráficamente los resultados y R contiene infinidad de funciones para crear gráficos de gran calidad.

En contrapartida, una de las desventajas de R es la lentitud en ejecutar operaciones computacionalmente complejas o que se tengan que repetir muchas veces. No obstante, es posible solventar este inconveniente escribiendo algunas funciones en lenguaje C++ y llamarlas desde R, de manera que el usuario trabaje exclusivamente desde R sin que "lo note" [42]. Escribir partes del algoritmo en C++ de forma adecuada puede acelerar muchísimo las operaciones llegando a ser más de 100 veces más rápido y haciendo factible el análisis de centenares de miles de variantes genéticas. Este hecho es de vital importancia, ya que, aunque el modelo estadístico diera estimaciones muy precisas, sería de nula utilidad para estudios con muchas variantes implicadas como son los GWAS.

OBJETIVOS

2.1. Objetivo general

El objetivo de esta tesis consiste en desarrollar modelos estadísticos y herramientas bioinformáticas eficientes para analizar estudios genéticos de asociación de variantes medidas con incertidumbre, tanto para estudios de casos y controles como para estudios de cohorte o transversales con respuesta continua. Para evaluar los modelos desarrollados, se han analizado tanto datos simulados bajo diferentes escenarios (variando la magnitud de la asociación, el grado de incertidumbre, etc.), como datos reales, y se han comparado los resultados con otros métodos ya existentes.

2.2. Objetivos específicos

1. Formulación del modelo propuesto para analizar la asociación entre las variantes genéticas con incertidumbre y una variable respuesta, en estudios caso-control y en estudios con rasgos cuantitativos.

2. Extensión del modelo a estudios de cohorte con datos censurados. Evaluación de la precisión y potencia estadística a la aplicación a SNPs imputados.

3. Extensión del modelo para el análisis de interacciones de variantes genéticas medidas con incertidumbre en estudios de cohorte y de caso control.

4. Implementación del algoritmo correspondiente al método propuesto al *software* libre `R` mediante la creación de un nuevo paquete de `R`, `CNVassoc`. Elaboración de una documentación extensa ilustrando el uso de las diferentes funciones (vignette). Integración de funciones escritas con lenguaje `C++` al paquete

`CNVassoc` a fin de acelerar notablemente los cálculos y hacer factible estudios donde se analizan centenares de miles de variantes (GWAS).

Para los objetivos (1 a 3) se han analizado tanto datos reales como datos simulados. Para el objetivo (1), los datos reales proceden de un *array* CGH, mientras que para los objetivos de (2) y (3), se han utilizado datos reales procedentes del estudio Framingham (http://www.framinghamheartstudy.org), y las simulaciones se han basado en estos datos, variando algunos parámetros como la magnitud de asociación, etc.

DISCUSIÓN DE LOS RESULTADOS Y CONCLUSIONES

Esta tesis se ha centrado en el análisis de datos obtenidos en estudios genéticos en los que las variantes genéticas se han medido con incertidumbre. En los cuatro artículos que se han publicado y que conforman esta tesis (ver capítulo 5) se ha propuesto un método para analizar la asociación entre variantes genéticas medidas con incertidumbre (CNVs o los SNPs imputados) para distintos tipos de estudios (de casos y controles, de cohorte o de respuesta cuantitativa). También se ha adaptado el modelo al análisis de interacciones de pares de SNPs imputados (epistasis). Todo ello se ha implementado en un paquete del *software* estadístico `R` llamado `CNVassoc` con un extenso manual donde se ilustran multitud de ejemplos (vignette). En el paquete `CNVassoc` se ha insertado código escrito en lenguaje `C++` para que los cálculos sean mucho más eficientes y así hacer factible el análisis de centenares de miles de variantes (GWAS).

A continuación se presentan y discuten los resultados más relevantes de esta tesis.

3.1. Incertidumbre en estudios de asociación

El principal problema que aparece en el análisis de datos genéticos medidos con incertidumbre es que si se usan modelos "clásicos" se induce un sesgo en la estimación de la asociación entre la variante genética y la respuesta. La razón de ello es que estos modelos suponen que todas las variables, tanto las explicativas como la respuesta, están medidas con total certeza en los individuos de la muestra. Éste no es

el caso que nos ocupa, sino que la variable explicativa de interés (el SNP imputado o el CNV) está medida con cierto grado de error en todos los individuos de la muestra.

Dependiendo del tipo de variable respuesta, el modelo que se ajusta es diferente: si la variable respuesta es continua y normalmente distribuida se construye un modelo de regresión lineal [43, 44], mientras que si la variable respuesta es binaria (presencia o ausencia de enfermedad, etc.) se suele usar un modelo de regresión logística [12, 45, 46], o si se trata de un estudio de cohorte con seguimiento, se ajusta un modelo de regresión de Cox [47, 48] o de Weibull [49].

El reto principal que intenta abordar esta tesis es cómo incorporar la incertidumbre observada en las variantes genéticas al ajustar modelos (regresión lineal, logística, de Cox, etc). En la literatura, se han descrito e implementado distintas alternativas, desde la más *naive* que supone que no hay incertidumbre y asigna el genotipo o número de copias más probable en cada individuo, hasta métodos mucho más complejos que estiman la probabilidad genotípica, a la vez que ajustan el modelo de asociación mediante modelos Bayesianos [39, 50, 51]. No obstante, esta tesis se ha centrado centrado en dos estrategias llamadas Naive y Dosage descartando las otras, básicamente por dos motivos: (i) estas dos estrategias son, sin ningún género de dudas, las más usadas e implementadas y (ii) son bastante simples y computacionalmente muy eficientes.

A pesar que hasta la fecha hay algunos estudios que analizan el comportamiento de estos dos métodos (Naive y Dosage), como por ejemplo Zheng *et al* en [52] para SNPs imputados, no se han encontrado en la literatura estudios de simulación suficientemente exhaustivos que comparen estos dos métodos con algún método de características similares al propuesto en esta tesis. Además, hay pocos estudios donde se hayan simulado datos bajo un diseño de cohorte, lo cual sí que se ha estudiado exhaustivamente a lo largo de los trabajos de esta tesis (**artículos 2 y 3**). Finalmente, otra aspecto novedoso de esta tesis es el análisis del comportamiento del método Naive y Dosage en el estudio de asociación de interacciones de SNPs imputados, poco o nada estudiados en la literatura, y se han comparado los resultados con los

obtenidos mediante el método propuesto.

3.2. Modelo propuesto

En los trabajos de esta tesis, se ha propuesto un modelo para el análisis de asociación de variantes genéticas medidas con incertidumbre. Este modelo se ha descrito formalmente y se ha planteado de forma explícita su verosimilitud (ver **artículo 1**). Su complejidad no es mucho mayor a la de los modelos clásicos de regresión donde las variables son medidas con certeza: se pueden ajustar utilizando métodos estándar de optimización como el procedimiento de *Newton-Raphson* (N-R) que está implementado en muchos *softwares*, también en el programa `R` [41].

Existen adaptaciones "numéricas" del método N-R, donde las primeras y segundas derivadas se aproximan de forma numérica dada la función de verosimilitud [53, 54, 55, 56]. No obstante, en este trabajo se han hallado de forma analítica todas sus derivadas, lo cual es de gran importancia en nuestra modelo porque se ha demostrado que acelera de forma "sorprendente" el proceso de optimización.

3.3. Aplicación a los CNVs

Para cada CNV se obtiene una intensidad de señal de una sonda prediseñada para detectar la variante en cuestión. Dependiendo del número de copias del segmento que tenga un individuo, esta intensidad será mayor o menor ya que se adhieren más o menos sondas a las cadenas de ADN. A partir de una muestra suficientemente grande de participantes, se obtiene una distribución de esta intensidad, que se espera que sea multimodal, donde cada "moda" corresponde a un grupo formado por todos los individuos con el mismo número de copias.

En el primer trabajo (**artículo 1**), se ha discutido cómo los métodos más "naive" aplicados para inferir el número de copias a partir de la intensidad de la sonda pro-

ducen resultados sesgados en el posterior análisis de asociación entre el CNV y la variable respuesta tanto en estudios de casos y controles como en estudios de respuesta binaria o respuesta continua. Estos métodos "naive" consisten en clasificar a los individuos a partir de puntos de corte sobre la distribución de la señal. Si la señal en cada grupo de individuos es muy homogénea y al mismo tiempo entre los grupos hay una separación clara, no habrá incertidumbre en clasificar a los individuos. Por el contrario, si estas distribuciones se solapan mucho, se pueden clasificar erróneamente un gran número de individuos, dando lugar a un sesgo en la asociación debido a que ésta queda diluida y por consiguiente se infraestima su efecto. Este hecho ya es bien conocido y se ha descrito en varios artículos, como en [36, 37], y también se ha visto en los resultados de este trabajo de la tesis.

Para cuantificar el grado de sesgo de los métodos existentes con el método propuesto, se han analizado datos simulados bajo distintos escenarios variando el efecto del CNV sobre la enfermedad y el grado de incertidumbre, o sea, el solapamiento de las distribuciones de la señal. Los resultados han sido muy concluyentes, siendo el modelo propuesto insesgado en todas las situaciones mientras que el método Naive infraestima mucho el efecto y es mucho menos potente.

La novedad del primer trabajo de la tesis ha sido la presentación e implementación de una técnica estadística que optimiza la verosimilitud, la cual se ha demostrado que es teóricamente la correcta en un modelo de asociación en que la variable explicativa (que en este caso es la variante genética) es medida con incertidumbre. Se ha visto también que el modelo propuesto converge en la gran mayoría de situaciones y computacionalmente no es mucho más costoso y que, por lo tanto, resulta factible analizar centenares e incluso miles de CNVs.

3.4. Aplicación a los SNPs imputados

Hasta la actualidad, la variante genética más estudiada con diferencia en los estudios de genética de poblaciones son los SNPs. Estas variantes son las más simples con lo que son más fáciles de genotipar y analizar. Gracias a la gran acumulación

de datos, ha sido posible estudiar patrones (LD) entre los SNPs, con los cuales es posible imputar o inferir un SNP no genotipado mediante otros SNPs cercanos o en gran LD a él. Ello ha permitido que los estudios de SNP a gran escala (GWAS) incorporen también SNPs no genotipados pasando de 1 millón de SNPs a 2 ó 3 millones más. Por consiguiente, es de vital importancia analizar de forma correcta los SNPs imputados en un estudio de asociación, ya que fácilmente pueden llegar a ser más de dos tercios del total de variantes estudiadas.

Como ocurre con los CNVs, no tener en cuenta la incertidumbre derivada de la imputación puede llevarnos a resultados sesgados y tener menos potencia para detectar SNPs asociados con la variable respuesta [37].

Existen varios tipos de algoritmos para imputar SNPs que pueden dar resultados distintos. No obstante, todos dan como *output* las probabilidades de cada uno de los tres genotipos para cada individuo de la muestra.

En el segundo trabajo de esta tesis (**artículo 2**) se ha demostrado como la aplicación del modelo propuesto inicialmente para CNVs se puede ampliar también a los SNPs imputados y que su funcionamiento es muy parecido a los CNVs: partiendo de las probabilidades imputadas para cada genotipo, es posible estimar el modelo de asociación optimizando la función de verosimilitud 1.1. Además, se ha comparado el método propuesto con un método Naive y con otro método ya implementado en muchos *softwares* y cómunmente usado en el análisis de SNPs imputados conocido como Dosage [37, 31]. Aunque existen otros métodos más complejos, estos dos métodos son con diferencia los más usados. El método Naive asigna el genotipo más probable a cada individuo y procede luego en el modelo de asociación como si no hubiera incertidumbre. El método Dosage, en cambio, sí que tiene en cuenta la incertidumbre, incorporando el número esperado de copias de cada individuo (Dosage) como variable predictora en el modelo [37]. Mientras que es conocido que el método Naive produce resultados sesgados y es poco potente, en teoría no está tan claro que el método Dosage asegure resultados insesgados y potentes, aunque ésto se afirme en [37]. Es importante, pues, no sólo comparar el modelo propuesto con las otras

dos estrategias (Naive y Dosage), sino que también es de igual importancia evaluar el comportamiento de estos dos métodos y en qué situaciones están sesgados, etc.

Hasta la fecha ningún estudio ha abordado la bondad del método Dosage. En uno de los artículos de esta tesis se analiza el posible sesgo y potencia estadística del método Dosage, bajo diversos escenarios que cubren un exhaustivo rango de posibilidades.

Los resultados obtenidos a partir de las simulaciones han demostrado que el método Dosage, en contra de lo comúnmente aceptado [37], es sesgado aunque este sesgo parece ser sólo importante en escenarios bastante extremos con efectos e incertidumbres de imputación grandes. El nuevo método propuesto en esta tesis se ha visto que se comporta de forma muy similar al Dosage, pero ha reportado resultados insesgados incluso en situaciones más extremas. Por último, el método Naive ha dado resultados muy sesgados y poco potentes en casi todos los escenarios, excepto cuando el grado de incertidumbre es casi nulo.

Todo ello, nos ha llevado a la conclusión de que el método Dosage se puede utilizar en casi todos los escenarios a menos que tanto el grado de incertidumbre con el que se ha imputado el SNP y/o efecto de de este SNP sean muy grandes, en cuyo caso es mejor usar el método propuesto. Mientras que el método Naive es mejor que sea evitado siempre que se analicen SNPs imputados. Por último, el método propuesto puede ser usado en cualquier escenario.

Des del punto de vista computacional, los métodos Dosage y Naive son equivalentes, dado que los dos ajustan modelos estándar, mientras que el método propuesto, al maximizar la función de verosimilitud 1.1 es más complejo. Aún así, en este trabajo se ha demostrado, después de computar los tiempos requeridos para el análisis de datos reales con centenares de miles de variantes, que es factible analizar un GWAS con el modelo propuesto y la diferencia de tiempo es imperceptible.

3.5. Aplicación a la interacción de SNPs imputados

Los estudios GWAS han tenido éxito en encontrar muchas variantes (SNPs) asociadas a distintas enfermedades o rasgos. Sin embargo, han dejado sin explicar gran parte de la variabilidad genética [57]. Se han buscado alternativas a los GWAS como son los estudios de otras variantes genéticas más complejas, por ejemplo los CNVs. Otros estudios han optado por averiguar si existen asociaciones en las interacciones de SNPs, con la premisa de que algunos SNPs por si solos no afectan al fenotipo estudiado o su efecto es demasiado pequeño, pero sí que interaccionan con otros SNPs y que el efecto de esa interacción es notable. Algunos de ellos han tenido éxito en encontrar interacciones asociadas con distintas enfermedades [58]. Con este objetivo, el tercer trabajo de esta tesis se ha centrado en el estudio de los modelos de asociación de interacciones de pares de SNPs imputados con cierta enfermedad (variable respuesta) en un estudio de cohorte y de casos y controles, denominados *Genome Wide Interaction Study* (GWIS).

Como ya se ha comentado anteriormente, hay muchos estudios alertando de la importancia de tener en cuenta la incertidumbre en la imputación de SNPs a la hora de estimar la asociación. Pero hay muy pocos artículos hasta la fecha que hayan analizado la influencia de esta incertidumbre en la interacción de SNPs imputados. Y es que un aspecto fundamental en los estudios de interacción (GWIS) es que la incertidumbre puede ser mucho más elevada que en los SNPs de forma individual (GWAS). Para los GWAS, a menudo se fija un punto de corte a partir del cuál se considera que la incertidumbre es suficientemente baja como para no tenerse en cuenta y en este caso se puede usar con "tranquilidad" el método Naive. Sin embargo, en los estudios de interacción, la incertidumbre tiene que ser evaluada por pares y no en cada SNP de forma individual. Así, una incertidumbre pequeña en los dos SNPs por separado, puede ser substancial en la interacción.

Para abordar el tema de la interacción, se ha supuesto que la incertidumbre en la

imputación del par de SNPs es independiente el uno del otro (no hay que confundir la independencia en la incertidumbre a la hora de imputar con el LD entre los dos SNPs). De esta forma, se puede calcular la probabilidad de cada par de SNPs (las nueve combinaciones) como el producto de las probabilidades. Se ha acomodado el modelo propuesto para el análisis de SNPs imputados o CNVs a interacciones de SNPs, y se ha aplicado a datos simulados bajo distintos escenarios variando el grado de incertidumbre de ambos SNPs, frecuencias alélicas y la magnitud de la asociación de la interacción. Además, al igual que para el estudio de SNPs "individuales", se han analizado datos reales procedentes del estudio de Framingham.

Los resultados obtenidos (**artículo 3**) han sido distintos de los obtenidos en los estudios de SNPs imputados "individuales": (i) el método Dosage empieza a ser sesgado con grados de incertidumbre más moderados, y este sesgo puede ser positivo o negativo dependiendo de las frecuencias alélicas, hecho que a priori no es nada intuitivo, y (ii) el método propuesto ha sido más potente que el método Dosage en escenarios de gran incertidumbre y efecto, a diferencia de lo mostrado anteriormente, cuando se analizaba un GWAS sin interacción y que tanto el método propuesto como el Dosage se mostraron casi equivalentes. Finalmente, al igual que en los GWAS, el método Naive ha sido muy sesgado y poco potente, si bien este hecho se ha manifestado con mayor grado si cabe que en los GWAS.

Otro hecho a destacar es que no se han encontrado programas donde se haya implementado el método Dosage para el estudio de GWIS con SNPs imputados. Este es un hecho remarcable ya que, (i) el método Dosage es muy usado para los GWAS, y, (ii) en este trabajo se ha demostrado que es trivial ampliar el método Dosage para GWAS a estudios de interacción de pares de SNPs imputados.

3.6. Implementación en un paquete de `R`

Cuando se diseña un modelo ó método estadístico para el análisis de datos, no sólo es importante que éste sea preciso y computacionalmente eficiente, sino que también es importante que pueda ser manejado de forma fácil para el máximo número

de usuarios. En el entorno de los análisis genéticos y también en el ámbito de la bioestadística, uno de los *softwares* más utilizado y conocido es `R`. El programa `R` contiene numerosas funciones y paquetes para el análisis de estudios genéticos de poblaciones, entre otros campos como son la econometría, la ecología, etc. Así pues, no es de extrañar que en esta tesis se haya optado por implementar e integrar las funciones y algoritmos correspondientes al modelo propuesto en `R`.

En el último y cuarto trabajo de esta tesis (**artículo 4**), se ha descrito con detalle la estructura de las funciones, objetos, clases, etc, juntamente con los documentos de ayuda debidamente escritos, que conforman el paquete llamado `CNVassoc` disponible en el repositorio CRAN de `R`. Además, se ha elaborado un manual (*vignette*) con numerosos y detallados ejemplos para que el usuario pueda reproducir y seguir, donde aparecen todas las funciones principales y sus opciones.

En el paquete `CNVassoc` se han construido funciones para analizar miles de SNPs a la vez. Estas funciones incorporan una opción para el análisis en paralelo a fin de acelerar el proceso. No obstante, se sabe que el código de `R` no es muy eficiente, así que también se ha escrito el algoritmo del modelo propuesto al lenguaje `C++` el cual es llamado desde la función principal de `R`. Con ello, la velocidad de ejecución del análisis se acelera notablemente y es factible analizar centenares de miles de SNPs en poco tiempo. El código en `C++` queda integrado dentro del paquete sin que el usuario "lo note'. Es decir, no tiene que escribir ninguna instrucción en `C++`, sólo en `R`. Más concretamente, el código traducido a `C++` fue el correspondiente al algoritmo de N-R para maximizar la función de verosimilitud del modelo propuesto con las primeras y segundas derivadas.

Otro reto a nivel computacional ha sido la simulación de los datos de cohorte en el segundo y tercer trabajo, donde la variable respuesta se ha generado a partir de la distribución empírica. Una vez más, debido a que el uso de código propio de `R` era muy ineficiente para generar la gran cantidad de datos que componían todas las simulaciones, se ha tenido que crear el código necesario en `C++`.

3.7. Conclusiones finales

1. Muchos de los estudios de asociación en epidemiología genética analizan variantes que no son medidas directamente sino que se infieren o imputan, dando lugar a incertidumbre. Este es el caso de los CNVs o de los SNPs imputados.

2. Las herramientas existentes hasta la fecha proponen modelos estadísticos más o menos fiables y precisos para estimar la asociación de las variantes genéticas con cierta enfermedad (variable repuesta). Se sabe que el hecho de no tener en cuenta la incertidumbre inherente en la medición de la variante genética conlleva resultados sesgados y poco potentes.

3. En esta tesis se ha propuesto e implementado un modelo que maximiza la verosimilitud teórica. Este modelo se ha formulado y comparado con otras técnicas comúnmente utilizadas hasta la fecha en los estudios de asociación genética para analizar CNVs por un lado, y SNPs imputados, por otro.

4. Hasta la fecha no se ha encontrado ningún estudio que discuta analíticamente porqué y hasta qué punto los métodos más utilizados son sesgados. Mediante exhaustivas simulaciones variando los escenarios en el grado de asociación e incertidumbre de las variantes genéticas, se ha visto que el método propuesto no es sesgado y tiene incluso más potencia estadística que los métodos existentes.

5. También se han analizado datos reales llegando a la conclusión que el modelo propuesto es útil para analizar estudios de GWAS con miles o centenares de miles de variantes y miles de individuos. Esto ha sido posible gracias a la implementación de parte del código al lenguaje `C++`, y al hecho de hallar de forma analítica las primeras y segundas derivadas en el proceso de optimización de la función de verosimilitud.

6. El modelo propuesto se ha extendido y adaptado a diferentes tipos de estudio, entre los cuáles, los de casos y controles y los de cohorte que son los más usuales en los estudios de asociación genética. Además, también se ha programado el modelo para soportar los análisis de interacciones de pares de variantes (epistasis), concretamente con SNPs imputados.

7. Finalmente, se ha creado un paquete integrado en el *software* `R` llamado `CNVassoc` donde se han ensamblado las diferentes funciones para el análisis de asociación en objetos y clases, a fin de que su uso sea familiar e intuitivo para los usuarios de este *software*. También se ha escrito un manual con múltiples ejemplos (*vignette*).

3.8. Trabajos futuros

3.8.1. Aplicaciones a otras variantes genéticas

En esta tesis se ha ilustrado cómo analizar datos genéticos medidos con incertidumbre para estimar su asociación con una variable respuesta (enfermedad/rasgo), y más concretamente para las variantes CNVs y SNPs imputados. Se ha propuesto un modelo para ello, el cual se ha demostrado que es válido y eficiente mediante estudios de simulación y análisis de datos reales.

Sin embargo, es fácil pensar que el modelo propuesto también podría aplicarse a otras variantes genéticas como son las inversiones, etc., siempre y cuando se verifiquen los siguientes requisitos: (a) que la variante genética sea categórica con un número limitado (a poder ser reducido) de categorías, (b) se haya medido con incertidumbre y que se obtenga, a partir de algún algoritmo, las distintas probabilidades de pertenecer a cada categoría para cada individuo de la muestra. El primer requisito haría referencia a los genotipos de los SNPs o al número de copias de los CNVs que suelen ser limitados (3 ó 4 como máximo), y el segundo sería el análogo a las probabilidades resultantes de los algoritmos de imputación de SNPs o de la inferencia de la señal de las sondas (mixturas de normales, etc) de los CNVs. Por otro lado, dados estos dos requisitos, y vistos los resultados satisfactorios obtenidos para los SNPs imputados y para los CNVs, también sería interesante aplicar el método Dosage a otras variantes genéticas y evaluar sus resultados.

3.8.1.1. *Next Generation Sequencing* (NGS)

La secuenciación de nueva generación, *Next Generation Sequencing* (NGS), es una técnica novedosa de genotipado de alto rendimiento capaz de determinar el orden de los nucleótidos de una cadena de ADN a alta densidad y a muy bajo coste, ya que es capaz de paralelizar un gran número de operaciones en la secuenciación [59, 60]. Existen diferentes herramientas bioinformáticas disponibles para analizar datos de NGS. En concreto, Bioconductor (http://www.bioconductor.org/) es un repositorio de `R` [41] donde se encuentran numerosos paquetes para tratar datos de este tipo, desde el análisis partiendo de datos crudos (normalización, lectura, pre-procesamiento, ...) hasta el análisis de datos pre-procesados (análisis estadístico, visualización, ...).

A *grosso modo* esta técnica replica segmentos de la cadena para posteriormente alinearlas con una de referencia para ese individuo. Dependiendo del número de réplicas y de la longitud de los segmentos se tiene más o menos cobertura. Una vez alineados los segmentos, se contabiliza la proporción de cada posible nucleótido ('A', 'T', 'C' ó 'G') en cada "locus". Es evidente que si sólo aparece un nucleótido se concluirá que en aquel locus el individuo es homozigoto. Por el contrario, si aparecen dos nucleótidos distintos se dirá que es heterozigoto. La ambigüedad viene dada cuando hay nucleótidos distintos en un mismo "locus" y estos porcentajes son por ejemplo 95 % y 5 %, ya que al poder existir cierto error en el alineamiento no queda claro que sea heterozigoto o monozigoto. Lo mismo ocurre cuando hay más de 2 nucleótidos distintos, en cuyo caso todavía es más difícil determinar el genotipo del individuo. Además, y de forma análoga a los SNPs imputados o a los CNVs, se define una medida del grado de incertidumbre que tiene que ver con la cobertura y la frecuencia alélica estimada para cada "locus".

En resumen, la técnica de NGS, aunque permite genotipar gran cantidad de ADN a un coste muy bajo (tiempo y dinero), tiene como contrapartida el hecho de que acarrea cierto error por lo explicado en el apartado anterior. Ello conlleva incertidumbre a la hora de determinar el genotipo. De igual forma que para los SNPs imputados o los CNVs es de vital importancia tener en cuenta esta incertidumbre en

estimar la asociación de las variantes genéticas determinadas por NGS y el fenotipo en cuestión, donde se podría aplicar el modelo propuesto en esta tesis adaptándolo a datos de NGS.

3.8.2. Aplicaciones a variantes no genéticas

Siguiendo los argumentos del apartado anterior, es lógico pensar que también se podrían analizar variables no genéticas con las mismas características que los SNPs imputados o los CNVs. Por ejemplo, en un estudio epidemiológico en que se desea estimar el efecto de la diabetes sobre el riesgo cardiovascular, posiblemente ajustando por otras variables. A veces, no es posible saber si un individuo es diabético o no, pero sí que se tienen otras variables relacionadas (los niveles de glucosa en sangre, etc) a partir de las cuales se puede inferir o estimar la probabilidad que cada individuo de la muestra sea diabético. En estos casos, se podría aplicar el modelo propuesto, donde el estatus de diabetes ocuparía el lugar de la variante genética (SNP imputado por ejemplo) y el hecho de ser diabético o no equivaldría al genotipo. Visto de esta manera, en este ejemplo, en lugar de haber 3 genotipos habría sólo 2.

Otro ejemplo interesante podría ser la aplicación del modelo propuesto como alternativa a las técnicas de imputación múltiple en el tratamiento de datos faltantes. Ello se podría hacer, lógicamente, siempre y cuando los datos faltantes estuvieran en una sola variable explicativa y ésta fuera categórica. En tal caso, los resultados obtenidos mediante imputaciones múltiples y usando el modelo propuesto serían equivalentes. La ventaja, pero, es que el modelo propuesto sería computacionalmente más eficiente que las técnicas de imputación múltiple, las cuales requieren simular unas cuantas veces la base de datos y ejecutar todos los análisis repetidamente (tantos como base de datos se hayan simulado).

FACTOR DE IMPACTO

Todas las publicaciones presentadas en esta tesis han sido publicadas en revistas internacionales especializadas en los campos de la bioinformática y genética. Las publicaciones, por orden en que se presentan en la tesis, son las siguientes:

Artículo 1:

Referencia	González JR, **Subirana I**, Escaramís G, Peraza S, Cáceres A, Estivill X, Armengol L. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. BMC Bioinformatics. 2009 Jun 6;10:172.
Número de citas	6
Número de accesos	3.389
Factor de impacto	**3,428**
Ranking	*Categoría: Total, Posición, Cuartil* - MATHEMATICAL & COMPUTATIONAL BIOLOGY: 29, 4, Q1 - BIOTECHNOLOGY & APPLIED MICROBIOLOGY: 152, 34, Q1 - BIOCHEMICAL RESEARCH METHODS: 67, 18, Q2

Artículo 2:

Referencia	**Subirana I**, González JR. Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies. Genet Epidemiol. 2013 Jul;37(5):465-77.
Número de citas	-
Número de accesos	No disponible
Factor de impacto	**4,015**
Ranking	*Categoría: Total, Posición, Cuartil* - PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH: 158, 14, Q1 - HEREDITY: 161, 43, Q2

Artículo 3:

Referencia	**Subirana I**, González JR. Interaction association analysis of imputed SNPs in case control and longitudinal studies. BMC bioinformatics. *Enviado.*
Número de citas	-
Número de accesos	-
Factor de impacto	**3,024**
Ranking	*Categoría: Total, Posición, Cuartil* - MATHEMATICAL & COMPUTATIONAL BIOLOGY: 47, 6, Q1 - BIOTECHNOLOGY & APPLIED MICROBIOLOGY: 159, 46, Q2 - BIOCHEMICAL RESEARCH METHODS: 75, 28, Q2

Artículo 4:

Referencia	**Subirana I**, Diaz-Uriarte R, Lucas G, González JR. CNVassoc: Association analysis of CNV data using R. BMC Med Genomics. 2011 May 24;4:47.
Número de citas	2
Número de accesos	4.252
Factor de impacto	**3,693**
Ranking	*Categoría: Total, Posición, Cuartil* - GENETICS & HEREDITY: 158, 44, Q2

Firmado por el director de la tesis:

Juan Ramón González Ruiz

Bibliografía

[1] L. Feuk, AR. Carson, and SW. Scherer. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97, 2006.

[2] BE. Stranger, MS. Forrest, M. Dunning, CE. Ingle, C. Beazley, N. Thorne, R. Redon, CP. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, SW. Scherer, S. Tavare, P. Deloukas, ME. Hurles, and ET. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[3] D. P. Locke, A. J. Sharp, S. A. McCarroll, S. D. McGrath, T. L. Newman, Z. Cheng, S. Schwartz, D. G. Albertson, D. Pinkel, D. M. Altshuler, and E. E. Eichler. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet*, 79(2):275–90, 2006.

[4] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.

[5] K. Wong, R. J. deLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng, D. E. Horsman, C. MacAulay, R. T. Ng, C. J. Brown, E. E. Eichler, and W. L. Lam. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet*, 80(1):91–104, 2007.

[6] M. Karayiorgou, C. Sobin, ML. Blundell, BL. Galke, L. Malinova, P. Goldberg, J. Ott, and JA. Gogos. Family-based association studies support a sexually dimorphic effect of comt and maoa on genetic susceptibility to obsessive-compulsive disorder. *Biol Psychiatry*, 45(9):1178–1189, May 1999.

[7] P. Sklar, SB. Gabriel, MG. McInnis, P. Bennett, Y-. Lim, G. Tsan, S. Schaffner, G. Kirov, I. Jones, M. Owen, N. Craddock, JR. DePaulo, and ES. Lander. Family-based association study of 76 candidate genes in bipolar disorder: Bdnf is a potential risk locus. brain-derived neutrophic factor. *Mol Psychiatry*, 7(6): 579–593, 2002.

[8] NM. Laird and C. Lange. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*, 7(5):385–394, May 2006.

[9] JN. Hirschhorn and MJ. Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108, Feb 2005.

[10] A. Rovelet-Lecrux, D. Hannequin, G. Raux, N. Le Meur, A. Laquerriere, A. Vital, C. Dumanchin, S. Feuillette, A. Brice, M. Vercelletto, F. Dubas, T. Frebourg, and D. Campion. App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*, 38 (1):24–6, 2006.

[11] J. Xing, RE. Myers, X. He, F. Qu, F. Zhou, X. Ma, T. Hyslop, G. Bao, S. Wan, H. Yang, and Z. Chen. GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis. *Eur J Cancer.*, 47(11):1699–707, Jul 2011.

[12] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet.*, 41(3):762, Mar 2009.

[13] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, R. J. O'Connell, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–40, 2005.

[14] C. Le Marechal, E. Masson, J. M. Chen, F. Morel, P. Ruszniewski, P. Levy, and C. Ferec. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet*, 38(12):1372–4, 2006.

[15] J. Ma, M. Xiong, M. You, G. Lozano, and CI. Amos. Genome-wide association tests of inversions with application to psoriasis. *Hum Genet*, [epub ahead of print], Mar 2014.

[16] C. Lippert, J. Listgarten, RI. Davidson, S. Baxter, H. Poon, CM. Kadie, and D. Heckerman. An exhaustive epistatic SNP association analysis on expanded wellcome trust data. *Sci Rep*, 3:1321, Feb 2013.

[17] SA. McCarroll and DM. Altshuler. Copy-number variation and association studies of human disease. *Nat Genet*, 39(7 Suppl):S37–42, 2007.

[18] International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[19] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7311):52–58, 2007.

[20] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[21] KL. Gunderson, FJ. Steemers, H. Ren, P. Ng, L. Zhou, C. Tsan, W. Chang, D. Bullis, C. Musmacker, J. anb King, LL. Lebruska, D. Barker, A. Oliphant, KM. Kuhn, and R. Shen. Whole-genome genotyping. *Methods Enzymol*, 410: 359–76, 2006.

[22] N. Rabbee and TP. Speed. A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, 22(1):7–12, Jan 2006.

[23] B. Fu and J. Xu. A new genotype calling method for Affymetrix SNP arrays. *J Bioinform Comput Biol*, 9(6):715–728, Dec 2011.

[24] CA. Anderson, G. Boucher, CW. Lees, A. Franke, M. D'Amato, KD. Taylor, JC. Lee, P. Goyette, M. Imielinski, A. Latiano, C. Lagacé, R. Scott, L. Amininejad, S. Bumpstead, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*, 43(3): 246–252, Mar 2011.

[25] PD. Pharoah, YY. Tsai, SJ. Ramus, CM. Phelan, EL. Goode, K. Lawrenson, M. Buckley, BL. Fridley, JP. Tyrer, H. Shen, R. Weber, R. Karevan, MC. Larson, H. Song, DC. Tessier, F. Bacot, D. Vincent, JM. Cunningham, J. Dennis, E. Dicks, Australian Cancer Study, Australian Ovarian Cancer Study Group, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*, 45(4):362–370, Apr 2013.

[26] K. Fellermann, D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, C. L. Bevins, W. Reinisch, A. Teml, M. Schwab, P. Lichter, B. Radlwimmer, and E. F. Stange. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon. *Am J Hum Genet*, 79(3):439–48, 2006.

[27] T. J. Aitman, R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A. J. Marshall, E. Petretto, M. D. Hodges, G. Bhangal, S. G. Patel, K. Sheehan-Rooney, M. Duda, P. R. Cook, D. J. Evans, J. Domin, J. Flint, J. J. Boyle, C. D. Pusey, and H. T. Cook. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078):851–5, 2006.

[28] AL. Price, NJ. Patterson, RM. Plenge, ME. Weinblatt, NA. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, Aug 2006.

[29] Y. Li, CJ. Willer, J. Ding, P. Scheet, and GR. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010.

[30] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[31] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, MAR. Ferreira, D. Bender, J. Maller, P. Sklar, PIW. de Bakker, MJ. Daly, and PC. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, 81(3):559–575, 2007. URL http://pngu.mgh.harvard.edu/purcell/plink/.

[32] J. Biernacka, R. Tang, J. Li, S. McDonnell, K. Rabe, J. Sinnwell, D. Rider, M. Andrade, E. Goode, and B. Fridley. Assessment of genotype imputation methods. *BMC Proceedings*, 3(Suppl 7):S5, December 2009.

[33] PI. de Bakker, MA. Ferreira, X. Jia, BM. Neale, S. Raychaudhuri, and Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17(R2):R122–8, Oct 2008.

[34] C. Fraley and A.E. Raftery. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *J Am Stat Assoc*, 97:611–631, 2002.

[35] Peter Macdonald and with contributions from Juan Du. *mixdist: Finite Mixture Distribution Models*, 2012. URL http://CRAN.R-project.org/package=mixdist. R package version 0.5-4.

[36] O. Davidov, D. Faraggi, and B. Reiser. Misclassification in logistic regression with discrete covariates. *Biometrical Journal*, 5:541â€“553, 2003.

[37] Y. Aulchenko, M. Struchalin, and C. van Duijn. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, 11:134, 2010.

[38] I. Ionita-Laza, AJ. Rogers, C. Lange, BA. Raby, and C. Lee. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 93(1):22–26, 2009.

[39] DY. Lin, Y. Hu, and BE. Huang. Simple and efficient analysis of disease association with missing genotype data. *Am J Hum Genet*, 82(2):444–452, 2008.

[40] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and ME. Hurles. A robust statistical method for case-control association testing with copy number variation. *Nat Genet*, 40(10):1245–1252, Oct 2008.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

[42] R Development Core Team. *Writing R extensions*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://CRAN.R-Project.org/doc/manuals/R-exts.html.

[43] VS. Voruganti, JW-Jr. Kent, S. Debnath, SA. Cole, K. Haack, HH. Göring, MA. Carless, JE. Curran, MP. Johnson, L. Almasy, TD. Dyer, JW. Maccluer, EK. Moses, HE. Abboud, MC. Mahaney, J. Blangero, and AG. Comuzzie. Genome-wide association analysis confirms and extends the association of SLC2A9 with serum uric acid levels to mexican americans. *Front Genet*, 4:279, Dec 2013.

[44] LV. Wain, GC. Verwoert, PF. O'Reilly, G. Shi, T. Johnson, AD. Johnson, M. Bochud, KM. Rice, P. Henneman, AV. Smith, GB. Ehret, N. Amin, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat Genet*, 43(10):1005–1011, Sep 2011.

[45] JS. Kooner, D. Saleheen, X. Sim, J. Sehmi, W. Zhang, P. Frossard, LF. Been, KS. Chia, AS. Dimas, N. Hassanali, T. Jafar, JB. Jowett, X. Li, V. Radha, SD. Rees, R. Takeuchi, F. andYoung, T. Aung, A. Basit, M. Chidambaram, D. Das, E. Grundberg, AK. Hedman, Islam M. HydrieZI, CC. Khor, S. Kowlessur, MM. Kristensen, S. Liju, WY. Lim, J. Matthews, DR. andLiu, AP. Morris, AC. Nica, JM. Pinidiyapathirage, I. Prokopenko, A. Rasheed, M. Samuel, N. Shah, AS. Shera, KS. Small, C. Suo, AR. Wickremasinghe, TY. Wong, M. Yang, DIAGRAM; MuTHER. ZhangF;, GR. Abecasis, AH. Barnett, M. Caulfield, P. Deloukas, Froguel P. FraylingTM, N. Kato, P. Katulanda, MA. Kelly, J. Liang, V. Mohan, J. Sanghera, DK. andScott, M. Seielstad, PZ. Zimmet, P. Elliott, YY. Teo, MI. McCarthy, J. Danesh, ES Tai, and JC. Chambers. Genome-wide association study in individuals of south asianancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet.*, 43(10):1984–9, Aug 2011.

[46] S. Trompet, A. de Craen, I. Postmus, I. Ford, N. Sattar, M. Caslake, D. Stott, B. Buckley, F. Sacks, J. Devlin, et al. Replication of LDL GWAs hits in PROSPER/PHASE as validation for future (pharmaco) genetic analyses. *BMC medical genetics*, 12(1):131, 2011.

[47] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genetics*, 5(3):e1000433, 2009.

[48] S. Bayraktar, PA. Thompson, SY. Yoo, KA. Do, AA. Sahin, BK. Arun, ML. Bondy, and AM. Brewster. The relationship between eight GWAS-identified single-nucleotide polymorphisms and primary breast cancer outcomes. *Oncologist*, 18(5):493–500, 2013.

[49] F. Del Greco M, C. Pattaro, A. Luchner, I. Pichler, T. Winkler, A.A. Hicks, C. Fuchsberger, A. Franke, S.A. Melville, A. Peters, et al. Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster. *Human molecular genetics*, 20(8):1660, 2011.

[50] Y. Guan and M. Stephens. Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12), Dec 2008.

[51] B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114, Jul 2007.

[52] J. Zheng, Y. Li, GR. Abecasis, and P. Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol*, 35 (2):102–110, Feb 2011.

[53] DF. Shanno and PC. Kettler. Optimal conditioning of quasi-Newton methods. *Math. Comp.*, 24:657–664, 1970.

[54] DF. Shanno. Conditioning of quasi-Newton methods for function minimization. *Math. Comp.*, 24:647–656, 1970.

[55] Goldfarb D. A family of variable-metric methods derived by variational means. *Math. Comp.*, 24:23–26, 1970.

[56] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

[57] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456 (7218):18–21, Nov 2008.

[58] Y. Kirino, G. Bertsias, Y. Ishigatsubo, N. Mizuki, I. Tugal-Tutkun, E. Seyahi, Y. Ozyazgan, FS. Sacli, B. Erer, H. Inoko, Z. Emrence, A. Cakar, N. Abaci, D. Ustek, C. Satorius, A. Ueda, M. Takeno, Y. Kim, GM. Wood, MJ. Ombrello, A. Meguro, A. Gül, EF. Remmers, and DL. Kastner. Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nat Genet*, 45(2):202–207, Feb 2013.

[59] N. Hall. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol*, 210(Pt 9):1518–25, May 2007.

[60] GM. Church. Genomes for all. *Sci Am*, 294(1):46–54, Jan 2006.

PUBLICACIONES

5.1. Copia de las publicaciones

- **Artículo 1.** *Accounting for uncertainty when assessing association between copy number and disease: a latent class model.* (pág. 54).

- **Artículo 2.** *Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies.* (pág. 67)

- **Artículo 3.** *Interaction association analysis of imputed SNPs in case control and longitudinal studies.* (pág. 99)

- **Artículo 4.** *CNVassoc: Association analysis of CNV data using R.* (pág. 80)

Nota: El artículo 3 se encuentra actualmente bajo revisión. Por consiguiente, se ha colocado en el apéndice de esta tesis.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Accounting for uncertainty when assessing association between copy number and disease: a latent class model

Juan R González*[1,2,3], Isaac Subirana[2,3], Geòrgia Escaramís[2,4], Solymar Peraza[2,1], Alejandro Cáceres[1,3], Xavier Estivill[4,2] and Lluís Armengol[4]

Address: [1]Center for research in environmental epidemiology (CREAL), Barcelona, Spain, [2]CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, [3]Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain and [4]Genes and Disease Program, Center for Genomic Regulation, Barcelona, Spain

E-mail: Juan R González* - jrgonzalez@creal.cat; Isaac Subirana - isubirana@imim.es; Geòrgia Escaramís - georgia.escaramis@crg.es; Solymar Peraza - speraza@creal.cat; Alejandro Cáceres - acaceres@creal.cat; Xavier Estivill - xavier.estivill@crg.es; Lluís Armengol - lluis.armengol@crg.es
*Corresponding author

Published: 06 June 2009

BMC Bioinformatics 2009, **10**:172 doi: 10.1186/1471-2105-10-172

Received: 12 November 2008
Accepted: 6 June 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/172

Abstract

Background: Copy number variations (CNVs) may play an important role in disease risk by altering dosage of genes and other regulatory elements, which may have functional and, ultimately, phenotypic consequences. Therefore, determining whether a CNV is associated or not with a given disease might be relevant in understanding the genesis and progression of human diseases. Current stage technology give CNV probe signal from which copy number status is inferred. Incorporating uncertainty of CNV calling in the statistical analysis is therefore a highly important aspect. In this paper, we present a framework for assessing association between CNVs and disease in case-control studies where uncertainty is taken into account. We also indicate how to use the model to analyze continuous traits and adjust for confounding covariates.

Results: Through simulation studies, we show that our method outperforms other simple methods based on inferring the underlying CNV and assessing association using regular tests that do not propagate call uncertainty. We apply the method to a real data set in a controlled MLPA experiment showing good results. The methodology is also extended to illustrate how to analyze aCGH data.

Conclusion: We demonstrate that our method is robust and achieves maximal theoretical power since it accommodates uncertainty when copy number status are inferred. We have made R functions freely available.

Background

With the recent technological advances, various genome-wide studies have uncovered an unprecedented number of structural variants throughout the human genome [1-3], mainly in the form of copy number variations (CNVs). The considerable number of genes and other regulatory elements that fall within these variable regions make CNVs very likely to have functional and, ultimately, phenotypic consequences [4,5]. In fact, recent studies have reported a correlation between copy number of specific genes and degree of disease predisposition [6-8], indicating that identification of DNA

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

copy number is important in understanding genesis and progression of human diseases.

Several techniques and platforms have been developed for genome-wide analysis of DNA copy number, such as array-based comparative genomic hybridization (aCGH). The goal of this approach is to identify contiguous DNA segments where copy number changes are present. The ability of aCGH to distinguish between different numbers of copies is limited, so various quantitative techniques are required for more precise, targeted analysis of genomic regions. For known CNVs, real time PCR assays can be used to compare the copy number status of particular loci in cases and controls. Individuals are typically binned into copy number categories using pre-defined thresholds of probe signal intensity. Recently, Multiplex Ligation-dependent Probe Amplification (MLPA) [9] has also been used to quantify copy number classes. This method allows the analysis of several loci at the same time in a single assay. MLPA is usually used to identify gains or losses in test samples with respect to controls [10], but it can also be used in the context of association studies in a case-control or cohort settings [11,12].

The statistical methods used in CNV-disease association studies are currently very simple. Quantitative methods give CNV probe signal intensity measurements for each individual as a continuous variable, from which copy number status is inferred, generally using pre-defined thresholds. Differences in copy number distribution between cases and controls are then assessed using χ^2, Fisher or Mann-Whitney tests [6,13,14]. However, the distribution of CNV probe measurements is continuous and multimodal, meaning that signal intensity should be considered as a mixture of curves. In many instances, these curves overlap with various underlying distributions leading to uncertainty. Therefore, scoring copy number by binning and then assessing the association may lead to misclassification and unreliable results.

Ionita-Laza et al. (2009) pointed out that it is not inmediately clear how this uncertainty of CNV calling should be incorporated in the statistical analysis [15]. To overcome this difficulty in assessing association between CNVs and disease, we propose a latent class (LC) model that incorporates possible uncertainty that appear when CNV calling is performed. After inferring copy number using Gaussian finite mixture distributions, or any other calling algorithm, the model assesses the relationship between the trait and a CNV using a mixture of generalized linear models. Association is then tested using a likelihood ratio procedure. We validate and compare our method with existing methods through a simulation study. We then illustrate how to test association between CNVs and the trait by using two real examples. One of them corresponds to a case-control study using data from a MLPA experiment where the true copy number status is known. The second example belongs to a study where breast cancer cell lines are analyzed using aCGH.

Methods

Inference of copy number status

Let us assume that we observe I individuals from a given population, consisting of $\mathcal{C}$ mutually exclusive latent classes c = 1, ..., $\mathcal{C}$ (e.g. copy number status). Instead of observing these classes, we observe a surrogate variable, X, corresponding to a continuous variable arising from any quantitative method. For instance, in targeted studies using MLPA or real-time PCR, X corresponds to peak intensities for each CNV probe. In the context of a whole genome scan, one may have quantitative data from aCGH or any other platform such as Illumina or Affymetrix, where, for each probe, the variable X corresponds to a ratio of intensities. Figure 1 shows a number of possible distributions that signal intensities may have. Some variants clearly show different underlying copy number status with multimodal signal intensities distributions (CNV2, CNV4 and CNV6). In other cases, where the existence of different copy numbers is not clear, inferring copy number by binning the data may be difficult or unfeasible.

For each CNV variant, we are interested in classifying the subjects into the $\mathcal{C}$ classes using the surrogate variable X. We propose to model the unobserved latent classes using a finite mixture model with $\mathcal{C}$ components of the form

$$f(x \mid \boldsymbol{\Theta}) = \sum_{c=1}^{\mathcal{C}} \pi_c N(x \mid \boldsymbol{\Theta}), \qquad (1)$$

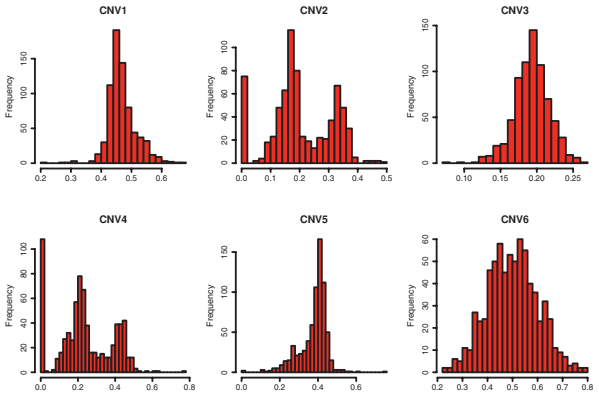


Figure 1
CNV quantitative measurements. Examples of CNV data showing different clustering quality and copy number status.

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

where $N(\cdot|\eta_c, \sigma_c^2)$ is the Gaussian distribution with Θ denoting all model parameters (e.g., $\Theta = (\eta_c, \sigma_c^2)$, $c = 1, ..., \mathcal{C}$), and x is the surrogate variable that corresponds to the quantitative measure of copy number status. For the component weights π_c it holds that

$$\sum_{c=1}^{\mathcal{C}} \pi_c = 1 \quad \text{and} \quad \pi_c \geq 0, \quad c = 1, ..., \mathcal{C}.$$

The value of $\mathcal{C}$ to be used is chosen by applying Bayesian Information Criteria (BIC) [16]. This mixture model approach for calling is similar to some used for the analysis of aCGH data [17,18] where correlation among probes should be considered. When analyzing MLPA data, it should be pointed out that in some instances, especially when there are individuals with 0 copies, the intensity distributions (see CNV2 and CNV4 in Figure 1) for a null allele is meant to be equal to 0. However, due to experimental noise it is fact that in some cases this ratio shows values that slightly deviate from this theoretical value. After our experience with hundreds of home-made MLPA probes, the value for null alleles is typically below 0.1; nevertheless, we recommend this parameter to be determined experimentally for each of the probes used in the MLPA experiments using the appropriate control samples. For these cases, the procedure used to estimate the parameters in (1) fails because the underlying distribution of individuals with 0 copies is not normal. In these situations we propose to fit the following mixture model to determine the latent classes

$$f(x \mid \Theta) = \pi_1 \mathcal{I}_{\{x \leq \tau\}} + \left(\sum_{c=2}^{\mathcal{C}} \pi_c N(x \mid \eta_c, \sigma_c^2) \right) \mathcal{I}_{\{x > \tau\}}, \quad (2)$$

where τ is given by the user, as previously indicated, $\pi_1 = \frac{\sum \mathcal{I}_{\{x \leq \tau\}}}{I}$, $\mathcal{I}$ denotes an indicator function, and

$$\pi_1 + \sum_{c=2}^{\mathcal{C}} \pi_2 = 1 \quad \text{and} \quad \pi_c \geq 0 \quad c = 2, ..., \mathcal{C}.$$

The posterior probabilities are used to segment data by assigning each individual to a given copy number status corresponding to the class with maximum posterior probability (MAP). After fitting this finite mixture model, we can perform a goodness-of-fit test using χ^2 test statistic. Finite mixture parameters can be estimated using the EM algorithm [19,20] or Newton-type procedures [20]. Then, the posterior probability that individual i with an observed value x belongs to copy number class j is given by

$$w_{ij} = \mathrm{P}(j \mid x, \Theta) = \frac{\pi_j N(x|\eta_j, \sigma_j^2)}{\sum_c \pi_c N(x|\eta_c, \sigma_c^2)}. \quad (3)$$

Latent class model

Discrete traits

Let us suppose that copy number status is associated with a binary phenotype (case-control). The association is typically assessed using a χ^2 test for the contingency table (Table 1). Misclassification in the table (due to uncertainty when inferring CNVs) is incorporated when we assign each individual to a given class c using *maximum a-posteriori probability* (MAP). Thus, this problem can be seen as an association study with misclas-sification ("measurement error") [21]. It is well known that misclassification of covariates has important implications for parameter estimates and statistical inference [22]. Some approaches account for such error [23,24]. These are, however, based on performing validation studies in a subsample. In the present context, this is unfeasible because hundreds of genes are normally analyzed at a time, and the technology may have a different sensitivity and specificity for each of the inspected loci. Therefore, we propose to use the posterior probability of belonging to each latent class to model the degree of misclassification of copy number status. We then take this information into account in the association model.

Conditioning on cluster c, we have that

$$\mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \boldsymbol{\beta}) = \mu_{ic}^{\gamma_i} (1 - \mu_{ic})^{1-\gamma_i}, \quad (4)$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_c)$, $c = 1, ..., \mathcal{C}$ is our vector of parameters, and

$$\mathrm{logit}(\mu_{ic}) \equiv \beta_c.$$

Then, equation (4) can be rewritten as

$$\mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \boldsymbol{\beta}) = \frac{e^{\gamma_i \beta_c}}{1 + e^{\beta_c}}.$$

Now, we consider that copy number status is measured with error (i.e., the latent class is not known). Therefore, we are modeling the probability of being an affected individual as a mixture of $\mathcal{C}$ binomial variables, as follows:

$$\mathrm{P}(\gamma_i \mid \boldsymbol{\beta}) = \sum_{c=1}^{\mathcal{C}} w_{ic} \mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \boldsymbol{\beta}),$$

Table 1: Contingency table of disease status and copy number category

	Copy number status				
Disease	1	2	...	$\mathcal{C}$	Total
Cases	r_1	r_2	...	$r_{\mathcal{C}}$	R
Controls	s_1	s_2	...	$s_{\mathcal{C}}$	S

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

where w_{ic} is the posterior probability that individual i belongs to copy number class c, given in (3). Therefore, assuming conditional independence of case-control status, given latent class, the likelihood function for model parameters $\boldsymbol{\beta}$ can be written as

$$\prod_{i=1}^{I}\sum_{c=1}^{\mathcal{C}} w_{ic}\mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \boldsymbol{\beta}) = \prod_{i=1}^{I}\sum_{c=1}^{\mathcal{C}} w_{ic}\frac{e^{\gamma_i\beta_c}}{1+e^{\beta_c}}. \quad (5)$$

We can then simply compute the odds ratio (OR) of belonging to class c with respect to a given reference r as

$$\mathrm{OR}_{c/r} = e^{\beta_c-\beta_r}. \quad (6)$$

Quantitative traits

We now consider the case where our phenotype, Y, is continuous. We assume that $Y \mid c\ N(\mu_c, \sigma^2)$. In this case, conditioning on cluster c

$$\mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \beta) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\gamma_i-\mu_{ic})^2}{2\sigma^2}}, \quad (7)$$

where

$$\mu_{ic} \equiv \beta_c.$$

Similar to the case of discrete traits, the likelihood function for model parameters $\boldsymbol{\beta}$ is given by

$$\prod_{i=1}^{I}\sum_{c=1}^{\mathcal{C}} w_{ic}\mathrm{P}(\gamma_i \mid \mathcal{C}_i = c, \beta) = \prod_{i=1}^{I}\sum_{c=1}^{\mathcal{C}} w_{ic}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\gamma_i-\beta_c)^2}{2\sigma^2}}. \quad (8)$$

In this case we are interested in evaluating the difference between the mean effect of individuals with c copies and r copies. This can simply be computed as

$$\overline{\gamma}_{c/r} = \beta_c - \beta_r.$$

Covariate Adjustment

In some instances researchers are interested in assessing the effect of CNVs after adjusting for other covariates, Z_1, ..., Z_K (usually called confounding variables). In this case, the likelihood function can be written as

$$\prod_{i=1}^{I}\sum_{c=1}^{\mathcal{C}} w_{ic}P(\gamma_i \mid \mathcal{C}_i = c, Z, \beta_c, \gamma),$$

where

$$P(\gamma_i \mid \mathcal{C}_i = c, Z, \beta_c, \gamma) = \frac{e^{\psi ic}}{1+e^{\psi ic}} \quad (9)$$

for discrete traits, and

$$P(\gamma_i \mid \mathcal{C}_i = c, Z, \beta_c, \gamma, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\gamma_i-\psi_{ic})^2}{2\sigma^2}} \quad (10)$$

for quantitative traits. In both cases

$$\psi_{ic} = \beta_c + \gamma_1 Z_{i1} + \ldots + \gamma_K Z_{iK}. \quad (11)$$

Parameter estimation

In this section we address parameter estimation for the general situation of having covariates and either discrete or quantitative traits. For brevity, let $\boldsymbol{\theta} \equiv (\beta, \gamma, \sigma)$ (notice that for discrete traits $\sigma = 1$). We consider that the weights, $\tilde{w}_{ic}$, are known and that they are given by the surrogate variable X from equation (3). Therefore, they can be used in the log-likelihood calculation, resulting in

$$\log P(\mathbf{Y} \mid \mathcal{C}_i = c, Z, \boldsymbol{\theta}) = \sum_{i=1}^{I}\log\sum_{c=1}^{\mathcal{C}}\tilde{w}_{ic}P(\gamma_i \mid \mathcal{C}_i = c, Z, \boldsymbol{\theta}). \quad (12)$$

Here $P(\gamma_i \mid \mathcal{C}_i = c, Z, \boldsymbol{\theta})$ is given by equations (9) and (10) for discrete and quantitative traits, respectively. The maximum likelihood estimators (MLE) of the model parameters maximize this log-likelihood function. We propose to use a Newton-Raphson procedure to find parameter estimates. The k-th component of the score, S, is given by

$$S_k(\gamma \mid \mathcal{C}, \boldsymbol{\theta}) \equiv \frac{\partial \log P(\mathbf{Y}|\boldsymbol{\theta})}{\partial\theta_k} = \sum_{i=1}^{I}\frac{\Sigma_{c=1}^{\mathcal{C}}\frac{\partial h_{ic}}{\partial\theta_k}}{\Sigma_{c=1}^{\mathcal{C}}h_{ic}}.$$

The k-th element of the Hessian, H, is

$$H_{kk'}(\boldsymbol{\theta}) \equiv \frac{\partial^2 \log P(\mathbf{Y}|\theta)}{\partial\theta_k\partial\theta_{k'}} = \sum_{i=1}^{I}\frac{\Sigma_{s=1}^{\mathcal{C}}\frac{\partial h_{ic}}{\partial\theta_k\partial\theta_{k'}}\Sigma_{s=1}^{\mathcal{C}}h_{ic} - \Sigma_{s=1}^{\mathcal{C}}\frac{\partial h_{ic}}{\partial\theta_k}\Sigma_{s=1}^{\mathcal{C}}\frac{\partial h_{ic}}{\partial\theta_{k'}}}{\left(\Sigma_{s=1}^{J}h_{ic}\right)^2}$$

where

$$h_{ic} \equiv \tilde{w}_{ic}P(\gamma_i \mid \mathcal{C}_i = c, Z, \theta).$$

Formulae for the derivatives of h_{ic} for covariates and for discrete and qualitative traits are given in the Appendix.

MLE can be used to estimate, under the multiplicative model, the OR between individuals with copy number

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

status c with respect to a reference category (e.g., individuals with copy number status r) as

$$\widehat{\mathrm{OR}}_{c/r} = e^{\hat{\beta}_c - \hat{\beta}_r}. \quad (13)$$

Similarly, when analyzing continuous traits, the estimated mean effect among individuals with c copies with respect to those with r copies is

$$\hat{\bar{\gamma}}_{c/r} = \hat{\beta}_c - \hat{\beta}_r. \quad (14)$$

The asymptotic variance-covariance matrix of maximum likelihood estimates of $\boldsymbol{\theta}$ can be estimated using the observed information matrix, F, as

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = F^{-1}(\hat{\boldsymbol{\theta}}) = -H^{-1}(\hat{\boldsymbol{\theta}}). \quad (15)$$

Therefore, we can compute a 95% confidence interval (CI95%) for $OR_{c/r}$ using the expression

$$CI_{1-\alpha}(\widehat{\mathrm{OR}}_{c/r}) \approx exp\left((\hat{\beta}_c - \hat{\beta}_r) \pm z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\boldsymbol{\theta}})_{[c,c]} + \widehat{Var}(\hat{\boldsymbol{\theta}})_{[r,r]} - 2\widehat{Var}(\hat{\boldsymbol{\theta}})_{[c,r]}} \right), \quad (16)$$

and for $\hat{\bar{\gamma}}_{c/r}$

$$CI_{1-\alpha}(\hat{\bar{\gamma}}_{c/r}) \approx (\hat{\beta}_c - \hat{\beta}_r) \pm z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\boldsymbol{\theta}})_{[c,c]} + \widehat{Var}(\hat{\boldsymbol{\theta}})_{[r,r]} - 2\widehat{Var}(\hat{\boldsymbol{\theta}})_{[c,r]}}, \quad (17)$$

where $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-th quantile of a standard normal distribution, α is the desired type-I error, and subindex $[\cdot, \cdot]$ denotes the position in the inverse of Fisher's information matrix.

Hypothesis testing

We propose to use a likelihood ratio test to assess disease association, taking the model without the copy number variable as reference. Twice the increase in the log-likelihood provides the asymptotic χ^2 statistic that tests H_0: $\beta_1 = \beta_2 = \ldots = \beta_{\mathcal{C}}$. In many instances, we are interested in studying the trend in effect with respect to copy number status (e.g., additive model). This can be done by generalizing equation (11) in the form

$$\psi_{ic} = \sum_{m=1}^{M} D_{icm}\zeta_{cm}, \quad (18)$$

where D is a $I \times M$ design matrix, and ζ is a vector of dimension M having the model parameters. M is the total number of variables included in the model, including copy number status and confounding variables (e.g., $M = \mathcal{C} + K$). For example, a trend test on copy number status without covariates D would have the form

$$\mathcal{D}' = \begin{pmatrix} 1 & \ldots & 1 & 1 & \ldots & 1 & \ldots & 1 & \ldots & 1 \\ 0 & \ldots & 0 & 1 & \ldots & 1 & \ldots & \mathcal{C}-1 & \ldots & \mathcal{C}-1 \end{pmatrix}$$

and the trend hypothesis on copy number status is tested using a likelihood ratio test, comparing this model with the null model. Notice that this formulation allows us to accommodate different or common effects for each latent class. In this case, parameter estimates are obtained as shown above. Formulae for the derivatives obtained in the score and Hessian, where coefficients are not shared by each latent class, are shown in the Appendix. R language functions for the methods discussed in this paper are freely available at http://www.creal.cat/jrgonzalez/software.htm [25]

Results

Simulation study

We performed computer simulation studies to empirically examine the properties of the parameter estimators developed in the previous sections. The specific goals of these studies were: (i) to evaluate the performance of the proposed likelihood ratio trend test based on the latent class model for a number of CNV measurement distributions; (ii) to examine the effect of sample size (I) on the distributional properties of the estimators; (iii) to examine the bias and mean square error (MSE) of the estimators; (iv) to assess the accuracy whether of the variance and parameter estimates obtained using the observed information matrix. Simulations were performed as follows: To study (i), we simulated a binary trait using 300 cases and 300 controls. The unobserved copy number statuses (e.g. latent classes) were simulated depending on 3 different copy number status ($\mathcal{C} = 3$), with the proportion of individuals in each category set as $\pi = (0.5, 0.4, 0.1)$. The trend OR was set equal to 1.5. The observed signal intensity ratio (X variable) were simulated as a finite mixture of $\mathcal{C}$ normal distributions using different means, η, and variances, σ^2, to assess whether the separation of clusters and their variance affects power.

To study (ii)–(iv) we simulated binary and quantitative traits. For the binary trait, simulation was performed as above but simulating various scenarios of sample size (I), OR and proportion of individuals with each copy number status, π. Again, we simulated different CNV distributions by varying η and σ^2. For quantitative traits, we used the same simulation procedure but copy number status was simulated depending on a fixed mean trait level for the reference copy number status and a desired mean difference with respect to other copy number statuses. Next, we describe the settings for the different simulation parameters. *Sample size*: We chose the values of I: $I \in \{50, 300\}$. Although current studies

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

are analyzing thousands of individuals, these values were chosen to evaluate the performance of our proposed method in moderately large samples. *Copy number status*: Since we were interested in evaluating the performance of the parameter estimates, we only simulated two different copy number statuses $\mathcal{C} = \{1, 2\}$. *Odds ratio*: To assess the impact of the strength of association between the disease and CNV, we chose two values for OR: OR $\in \{1.3, 2\}$ in order to consider a moderate association and a strong one. *Proportion of cases with normal copy number status*: To evaluate the impact of classes with different number of individuals we set $\pi \in \{(0.8, 0.2), (0.5, 0.5)\}$. *Finite mixture*: To asses the impact of distribution of intensity ratio, X, we simulated two normal distributions with the following parameters: $\eta \in \{1, 1.5\}$, which correspond to having 2 (considered as normal copy number status) and 3 copies, respectively, and $\sigma \in \{(0.15, 0.15), (0.15, 0.2), (0.2, 0.2)\}$. In this case, these scenarios also helped us to model different situations regarding misclassification or how latent classes were separated.

We compared three different approaches. The first (NAIVE) was based on assessing association between disease and copy number status obtained using MAP from the finite mixture model (2). That is, association was assessed using a χ^2 test from Table 1. The second is the approach that has been used predominantly to date when analyzing this kind of data and is based on assigning CNV status using pre-defined thresholds (THRES). Association is then assessed using a χ^2 test. As mentioned previously, we simulated data from two mixtures of normal distributions with means of 1 and 1.5. This is equivalent to simulating individuals with 2 and 3 copies, respectively. In this situation, it is considered that individuals with intensity (or intensity-ratio) greater than 1.33 correspond to individuals with 3 copies [10]. The third method is the one proposed in this paper, based on latent class (LC) using a χ^2 test. In order to make the results comparable, the performance of LC based on likelihood ratio trend test was compared with that of the two other methods using a χ^2 trend test (e.g. 1 degree of freedom). To evaluate bias and MSE of parameter estimates, χ^2 of association was used for all three methods.

Simulation results for evaluating the performance of the likelihood ratio trend test in our proposed model are shown in Figure 2. The top figures show the power for all methods analyzed under two scenarios (other scenarios are given in Additional file 1).

The left panel shows the power for each method, varying the CNV measurement distribution with regard to the mean of each latent class, η, while the right panel gives the same information but with fixed means and varying variances, σ^2. Figure 2 also depicts the distribution of CNV signal intensities for various scenarios. We observe that our proposed latent class model performs better in all cases, even when distribution of copy number status are not very well separated (e.g. more uncertainty).

Simulation results to evaluate parameter estimates for discrete traits are presented in Table 2 and in Table S1 and Figures S3 and S4 (see Additional file 1). Similar results and conclusions are obtained for a quantitative trait. Table 2 and Figures S3 and S4 (see Additional file 1) summarize the OR obtained by comparing individuals with 3 copies to those with 2 copies (reference category) and give the MSE for two different sample sizes, I, two different proportions of individuals with 2 copies, π, and two different variances for each component of the mixture, σ. Table S1 (see Additional file 1) compares different methods to compute

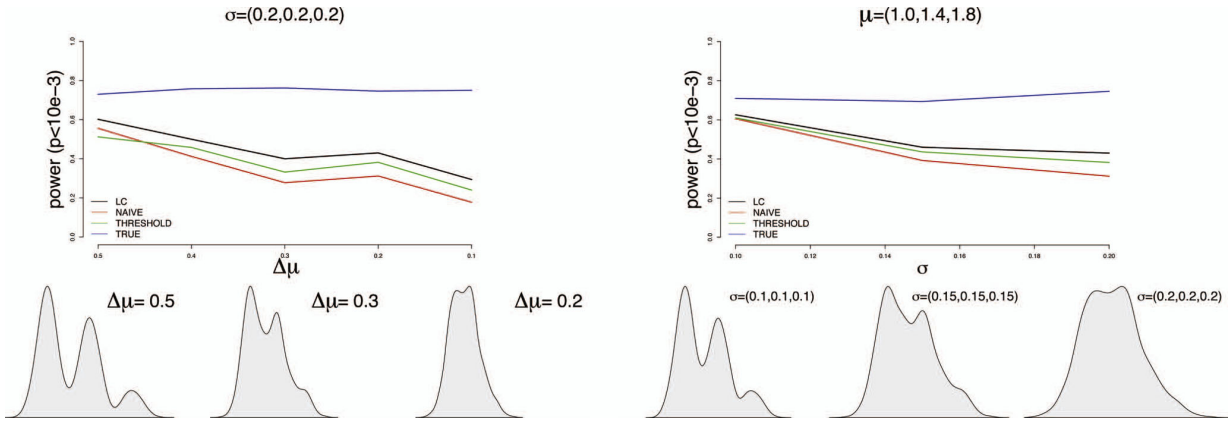


Figure 2
Empirical power for simulation studies. Empirical power for the three different approaches analyzed, varying the quality of clustering for underlying copy number status. Left panel is for fixed variance and varying means, while the right panel is for fixed mean and varying variances.

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

Table 2: Simulation study

				$e^{\hat{\beta}}$				Mean Square Error ($\times 10^3$)		
I	π	e^{β}	σ	SIM	NAIVE	THRES	LC	NAIVE	THRES	LC
50	0.8	1.3	(0.15,0.15)	1.23	1.17	1.15	1.20	57	87	42
50	0.8	1.3	(0.2,0.2)	1.24	1.14	1.09	1.21	107	131	114
50	0.8	1.3	(0.15,0.2)	1.28	1.18	1.15	1.24	134	148	112
50	0.8	2	(0.15,0.15)	1.60	1.40	1.28	1.48	54	85	44
50	0.8	2	(0.2,0.2)	1.82	1.36	1.29	1.52	152	158	126
50	0.8	2	(0.15,0.2)	1.89	1.42	1.33	1.57	180	253	162
50	0.5	1.3	(0.15,0.15)	1.26	1.24	1.21	1.26	39	51	32
50	0.5	1.3	(0.2,0.2)	1.32	1.28	1.25	1.35	82	79	97
50	0.5	1.3	(0.15,0.2)	1.26	1.23	1.20	1.26	66	72	60
50	0.5	2	(0.15,0.15)	2.04	1.94	1.83	2.05	40	67	34
50	0.5	2	(0.2,0.2)	2.04	1.76	1.68	2.05	107	128	92
50	0.5	2	(0.15,0.2)	2.06	1.78	1.72	1.99	87	107	71
300	0.8	1.3	(0.15,0.15)	1.30	1.25	1.18	1.30	13	32	10
300	0.8	1.3	(0.2,0.2)	1.32	1.25	1.15	1.34	27	50	29
300	0.8	1.3	(0.15,0.2)	1.30	1.22	1.16	1.29	24	42	21
300	0.8	2	(0.15,0.15)	2.01	1.87	1.49	2.01	21	120	13
300	0.8	2	(0.2,0.2)	2.03	1.70	1.36	1.99	69	203	43
300	0.8	2	(0.15,0.2)	2.03	1.62	1.38	1.86	78	189	38
300	0.5	1.3	(0.15,0.15)	1.31	1.27	1.26	1.30	7	9	5
300	0.5	1.3	(0.2,0.2)	1.30	1.23	1.22	1.30	15	17	12
300	0.5	1.3	(0.15,0.2)	1.30	1.24	1.23	1.29	12	14	9
300	0.5	2	(0.15,0.15)	2.00	1.87	1.77	2.00	11	23	5
300	0.5	2	(0.2,0.2)	2.00	1.72	1.66	2.02	36	51	15
300	0.5	2	(0.15,0.2)	2.00	1.76	1.71	1.97	26	37	10

Odds ratio (e^{β}) and mean square error obtained in 1,000 simulations using the three different approaches, NAIVE, THRES and LC (see text for a description of each). Results are given for different scenarios, varying the number of individuals (I), the proportion of individuals with each copy number status (π), the odds ratio (e^{β}), and the variance for CNV quantitative measurements.

the standard error of the ORs for the various scenarios described above. The results compare asymptotic variance based on an observed information matrix (ASYM) with respect to empirical variance (EMP). Supplementary Table S1 also shows coverage and power of confidence intervals based on the three methods analyzed. As expected, when the sample size increased, the performance of the estimators of the finite-dimensional parameters improved (Table 2). In all cases, the LC method performs better than the others. LC has less bias than NAIVE and THRES in all cases, and also shows better MSE.

Regarding variance estimates, the estimation based on ASYM showed good performance in all scenarios (see Additional file 1, Table S1). Despite slightly overestimating of EMP, the bias was less pronounced for I = 300, as expected. Confidence intervals based on the LC method outperform those obtained by other methods with regard to power.

Application to real data

MLPA example

The first data set used to analyze CNV and disease was generated and kindly provided by one of the coauthors of the current work. Although data is still unpublished, it has been made available in a blinded format for reproducing our findings using the approach presented herein, and for other validation studies. Some candidate genes were identified after performing a whole genome scan analysis using aCGH, where a pool of controls and cases were compared. In order to further investigate the relationship between the disease and altered the genes, a targeted study including several variants was designed using the MLPA technique. We obtained signal intensities of MLPA assays for 360 cases and 291 controls. Figures 3 and 4 show the intensities for cases and controls for two selected genes. In both cases, we observe 3 latent classes, corresponding to 0, 1, and 2 copies of the gene. We found that the finite mixture model fits very well (χ^2 goodness-of-fit test, P = 0.6615 and P = 0.4888). The main difference between these two cases is that copy number status for gene 1 can be established using a threshold method, while for the second gene this classification seems more arbitrary. As a consequence, misclassification should be taken into account when analyzing gene 2. Table 3 shows the classification of individuals as having 0, 1, 2 copies, estimated using equation (2) and the true copy number obtained by breakpoint cloning and assessing allele presence by PCR, which unequivocally reports the exact number of copies.

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

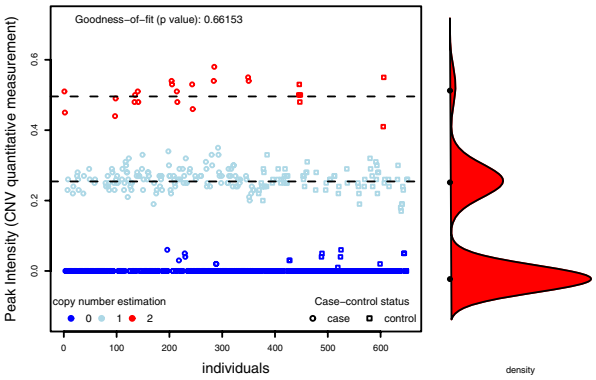


Figure 3
Association between Gene 1 and disease. Graphical representation of peak intensities (CNV quantitative measurement) of individuals for Gene 1 analyzed in the example. The various colors indicate copy number status inferred using our proposed finite mixture model.

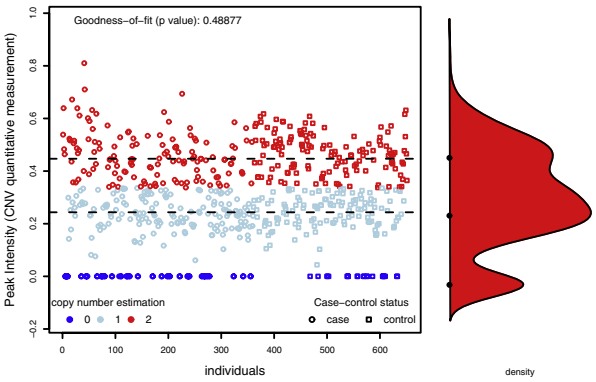


Figure 4
Association between Gene 2 and disease. Graphical representation of peak intensities (CNV quantitative measurement) of individuals for Gene 2 analyzed in the example. The various colors indicate copy number status inferred using our proposed finite mixture model.

From the table, we can see that the finite mixture model gives a perfect classification for gene 1 and some misclassification for gene 2. Goodness-of-fit test revealed that the proposed mixture model to determine CNV status was appropriate (p = 0.6615 and p = 0.1586).

Table 4 shows the ORs and their 95%CI for the two genes analyzed. The first three columns show the results obtained in the laboratory using PCR, while the other columns show the results obtained after estimating the copy number status using our proposed finite mixture model and computing the ORs using a naïve approach (e.g. assuming that there is no misclassification) and the LC model that accounts for misclassification. As we can see, the results are the same for gene 1, since no misclassification is observed (see Figure 3 and Table 3). However, for gene 2, copy number status could not be determined as easily as for gene 1. Thus, we observe a different OR estimation and, more importantly, a different *P*-value for association. For instance, the order of magnitude of the association between the disease and gene 2 is better captured by the LC model than by the NAIVE approach. Regarding the OR estimates, the analysis using the true copy number status shows that individuals with one copy of gene 2 have a 63% decrease in disease risk with respect to individuals with 0 copies. As the 95%CI shows, this difference is statistically significant. We arrive at the same conclusion when we compare individuals with 2 copies with respect to those with 0 copies. Note that in both cases we observe that the naïve approach underestimates the OR, as shown by the simulation study.

Table 3: Contingency table of estimated and true copy number status for the two genes examined in the real data example

	True copy number status		
	0	1	2
Gene 1			
0	426	0	0
1	0	201	0
2	0	0	24
Gene 2			
0	85	0	0
1	5	287	0
2	0	73	204

aCGH example

The analysis of aCGH data requires additional steps to take into account the dependency across probes. Table 5 shows four steps we recommend for the analysis of this kind of data. First, MAP should be obtained with an algorithm that considers probe correlation. We use, in particular, the CGHcall R program which includes a mixture model to infer CNV status [18]. Second, we build blocks/regions of consecutive clones with similar signatures. To perform this step the CGHregions R library was used [26]. Third, the association between the CNV status of blocks and the trait is assessed by incorporating the uncertainty probabilities in the LC model. And fourth, corrections for multiple comparisons must be performed. We use the Benjamini-Hochberg (BH) correction [27]. This is a heuristic method that is robust against positive dependence and increasingly conservative as correlation increases [28].

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

Table 4: Association analysis of disease status and copy number category using the true copy number status and the estimated status obtained using the finite mixture proposed

	True CN			Estimated CN			
	Co	Ca	OR (CI95%)	Co	Ca	$OR_{naïve}$ (CI95%)	OR_{LC} (CI95%)
Gene 1							
0	210	216	1	210	216	1	1
1	75	126	1.63 (1.16,2.30)	75	126	1.63 (1.16,2.30)	1.63 (1.16,2.30)
2	6	18	2.92 (1.14,7.49)	6	18	2.92 (1.14,7.49)	2.92 (1.14,7.50)
P association			0.0027			0.0027	0.0023
P trend			5.0×10^{-4}			5.0×10^{-4}	5.0×10^{-4}
Gene 2							
0	24	66	1	22	63	1	1
1	159	201	0.46 (0.27,0.77)	129	178	0.44 (0.26,0.75)	0.47 (0.27,0.82)
2	108	93	0.31 (0.18,0.54)	140	119	0.33 (0.19,0.57)	0.31 (0.18,0.54)
P association			7.2×10^{-5}			2.3×10^{-4}	8.4×10^{-5}
P trend			2.1×10^{-5}			1.0×10^{-4}	2.1×10^{-5}

Table 5: Steps used to assess association between CNVs and traits when aCGH is used

Step 1. Use any aCGH calling procedure that provides MAP (uncertainty)
Step 2. Build blocks/regions of consecutive probes with similar signatures
Step 3. Use the signature that occurs most in a block to perform association unsing LC model
Step 4. Correct for multiple testing considering dependency among signatures

We applied the methodology to the breasts cancer data studied by Neve et al. [29], which is freely available from the bioconductor website http://www.bioconductor.org/ [30]. The data consists on CGH arrays of 1 MB resolution [31]. The authors chose the 50 samples that could be matched to the name tokens of caArrayDB data (June 9th 2007).

In this example the association between strogen receptor positivity (dichotomous variable; 0: negative, 1: positive) and CNVs was tested. We contrasted the association as given by the LC and the NAIVE models. The original data set contained 2621 probes which were reduced to 459 blocks after the application of CGHcall and CGHregions functions. Table 6 shows the number of CNV blocks associated with strogen receptor positivity for different

Table 6: Number of CNV blocks (out of 459) associated with estrogen receptor positivity from 50 aCGH breast cancer cell lines

	Significance level				
	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
Latent class model	1	4	27	64	117
Chi-square test	0	2	10	41	93

Results are given for different levels of association and comparing our proposed model with the naïve approach that does not consider uncertainty.

significance levels. We observe that incorporating classification uncertainty with the LC model substantially increased the level of association, as compared to the NAIVE approach. The number of positive association at 5% of significance after applying BH correction was 49 and 24 for LC and NAIVE approach, respectively.

Discussion

In this paper we have shown that the assessment of association between CNVs and disease using analysis methods that do no take into account uncertainty when inferring copy number status lead to larger p-values and underestimate the model parameters. This confounds the need to increase statistical power, which is reduced by the multiple comparison correction for the simultaneous testing of several loci. False positives are typically controlled by a dramatic reduction in the nominal p-value, such that very low values are required to reach statistical significance. Thus, a precise computation of these values is essential in genetic association studies.

Here we have proposed a latent class model (LC) that accounts for the uncertainty of assessing CNV status and also accommodates potential confounding factors. In the case of analyzing quantitative traits, we also provide formulae to further propagate call uncertainty, as other authors have proposed in another context [32]. By analyzing quantitative traits, we have assumed that the response variable follows a normal distribution, although this assumption does not hold in some instances. In this situation, one possibility is to analyze the log-transformed variable, although log transformation may not be not sufficient. The model could easily be extended to fit a response variable that has any exponential family distribution (e.g. normal, gamma, Poisson). However, we have not yet implemented this option in the functions reported here. The extension of our proposed latent-class

model to assess survival time, possibly with right-censored data, is not trivial but could be a very interesting avenue for future investigation. The parameter estimation procedure proposed here, allows the estimation of confidence intervals. The LC model was remarkably consistent with simulated data. In particular, we found that the p-values obtained with the LC model were more similar to the expected values than those obtained by the threshold and naïve methods.

We maximize the likelihood function, assuming fixed weights for each copy number status, which accounts for possible misclassification. The main advantage of considering weights as known constants is that the Newton-Raphson procedure is much simpler, faster and feasible for obtaining the Hessian matrix analytically. We confirmed that the proposed model captures very well the nature of the synthetic data and variance estimates. Interestingly, we observed that the variance estimates using MLE were also reproduced when a bootstrap procedure was used (see Additional file 1, Table S2). In the interest of generalization, one can consider maximizing the likelihood function for both model parameters and weights. In that case, an EM algorithm should be used instead. However, one should bear in mind that EM does not allow for estimation of the variance of the model parameters and is computationally expensive, which may be particularly costly if this method is used in whole genome scan settings.

Conclusion

We have shown that the LC model can incorporate uncertainty of CNV calling in the analysis. We have also illustrated how to analyze quantitative traits as well as how to accomodate confounding variables. This is of particular importance in complex diseases studies where other clinical or biochemical factors need to be taken into account. The formulation can also be generalized to assess survival times or counts in longitudinal studies. The model has showed good performance when analyzing both targeted (MLPA data) and whole genome (aCGH data) studies.

Authors' contributions

JRG and IS developed the new statistical methods. JRG wrote the R functions and the main text of the manuscript and performed the simulation studies. GE and AC made abundant suggestions for developing the models. SP worked on the gaussian mixture approach to model quantitative CNVs measurements. XE reviewed the paper and revised its framework. LA and JRG proposed the need of a statistical tool to measure the biological differences in allele distribution in cohorts of cases and controls, and conceived the study. All authors have read, and approved the final manuscript.

Appendix

To obtain parameter estimates, we maximize the log-likelihood function

$$\log P(\boldsymbol{Y} \mid \mathcal{C}_i = c, Z, \boldsymbol{\theta}) = \sum_{i=1}^{I} \log \sum_{c=1}^{\mathcal{C}} \tilde{w}_{ic} P(\gamma_i \mid \mathcal{C}_i = c, Z, \boldsymbol{\theta}),$$

where $P(\gamma_i | \mathcal{C}_i = c, Z, \boldsymbol{\theta})$ is given by equations (9) and (10) for discrete and quantitative traits, respectively. As previously mentioned, the k-th component of the score, S, is given by

$$S_k(\gamma \mid \mathcal{C}, \boldsymbol{\theta}) \equiv \frac{\partial \log P(\boldsymbol{Y}|\boldsymbol{\theta})}{\partial \theta_k} = \sum_{i=1}^{I} \frac{\sum_{c=1}^{\mathcal{C}} \frac{\partial h_{ic}}{\partial \theta_k}}{\sum_{c=1}^{\mathcal{C}} h_{ic}}.$$

The k-th element of the Hessian, H, is

$$H_{kk'}(\boldsymbol{\theta}) \equiv \frac{\partial^2 \log P(\boldsymbol{Y}|\boldsymbol{\theta})}{\partial \theta_k \partial \theta_{k'}} = \sum_{i=1}^{I} \frac{\sum_{s=1}^{\mathcal{C}} \frac{\partial h_{ic}}{\partial \theta_k \partial \theta_{k'}} \sum_{s=1}^{\mathcal{C}} h_{ic} - \sum_{s=1}^{\mathcal{C}} \frac{\partial h_{ic}}{\partial \theta_k} \sum_{s=1}^{\mathcal{C}} \frac{\partial h_{ic}}{\partial \theta_{k'}}}{\left(\sum_{s=1}^{I} h_{ic}\right)^2}$$

where

$$h_{ic} \equiv \tilde{w}_{ic} P(\gamma_i \mid \mathcal{C}_i = c, Z, \theta).$$

Herein we provide formulae for the derivatives of h_{ic} for all cases discussed in this paper. Although the following expressions may appear complicated, they are straightforward to program and are included in the >R functions available at http://www.creal.cat/jrgonzalez/software.htm.

Binary Traits

Binary Traits without covariates

In this case, the h_{ic} function takes the form

$$w_{ic} \frac{e^{\gamma_i \beta_c}}{1+e^{\beta_c}}.$$

Therefore,

$$\frac{\partial h_{ic}}{\partial \beta_k} = \frac{w_{ic} \mathcal{I}_{\{k=c\}} \gamma_i e^{\gamma \beta_k} (1+e^{\beta_k}) - a_{ic} \mathcal{I}_{\{k=c\}} e^{\gamma_i \beta_k} e^{\beta_k}}{(1+e^{\beta_k})^2} = \mathcal{I}_{\{k=c\}} h_{ic} (\gamma_i - p_{ic}),$$

where

$$p_{ic} = \frac{1}{1+e^{-\beta_c}},$$

and

$$\frac{\partial^2 h_{ic}}{\partial \beta_k^2} = \mathcal{I}_{\{k=c\}} \left[\frac{\partial h_{ic}}{\partial \beta_k} (\gamma_i - p_{ic}) - h_{ic} (p_{ic} - p_{ic}^2) \right], \quad \text{and} \quad \frac{\partial^2 h_{ic}}{\partial \beta_k \beta_{k'}} = 0 \text{ for } k \neq k'.$$

BMC Bioinformatics 2009, **10**:172 http://www.biomedcentral.com/1471-2105/10/172

Binary Traits with covariates

In this case, the h_{ic} function takes the form

$$h_{ic} = w_{ic}\frac{e^{\gamma_i\psi ic}}{1+e^{\psi ic}}, \quad \text{where} \quad \psi_{ic} = \beta_c + \sum_{k=1}^{K}\gamma_k z_{ik}.$$

Therefore,

$$\frac{\partial h_{ic}}{\partial \beta_k} = \frac{w_{ic}\mathcal{I}_{\{k=c\}}\gamma_i e^{\gamma\psi ic}(1+e^{\psi ic}) - w_{ic}\mathcal{I}_{\{k=c\}}e^{\gamma_i\psi ic}e^{\psi ic}}{(1+e^{\psi ic})^2} = \mathcal{I}_{\{k=c\}}h_{ic}(\gamma_i - p_{ic}),$$

where

$$p_{ic} = \frac{1}{1+e^{-\psi ic}},$$

and

$$\frac{\partial^2 h_{ic}}{\partial \beta_k^2} = \mathcal{I}_{\{k=c\}}\left[\frac{\partial h_{ic}}{\partial \beta_k}(\gamma_i - p_{ic}) - h_{ik}(p_{ic} - p_{ic}^2)\right], \quad \text{and} \quad \frac{\partial^2 h_{ic}}{\partial \beta j \partial \beta j'} = 0 \text{ for } k \neq k'.$$

For covariates:

$$\frac{\partial h_{ic}}{\partial \gamma_p} = z_p h_{ic}(\gamma_i - p_{ic})$$

$$\frac{\partial^2 h_{ic}}{\partial \gamma_p^2} = z_p\frac{\partial h_{ic}}{\partial \gamma_p}(\gamma_i - p_{ic}) - z_p^2 h_{ic}(p_{ic} - p_{ic}^2)$$

$$\frac{\partial^2 h_{ic}}{\partial \gamma_p \partial \gamma_{p'}} = z_p\frac{\partial h_{ic}}{\partial \gamma_{p'}}(\gamma_i - p_{ic}) - z_p z_{p'} h_{ic}(p_{ic} - p_{ic}^2)$$

Quantitative traits

Quantitative traits without covariates and shared variance

In this case, the h_{ic} function takes the form

$$h_{ic} = w_{ic}\frac{1}{\sigma}e^{-\frac{(\gamma_i-\beta_c)^2}{2\sigma^2}}.$$

Therefore,

$$\frac{\partial h_{ic}}{\partial \beta_k} = \mathcal{I}_{\{k=c\}}w_{ic}\frac{1}{\sigma}e^{-\frac{(\gamma_i-\beta_k)^2}{2\sigma^2}}\frac{\gamma_i-\beta_k}{\sigma^2} = \mathcal{I}_{\{k=c\}}h_{ic}\frac{\gamma_i-\beta_k}{\sigma^2}$$

$$\frac{\partial^2 h_{ic}}{\partial \beta_k^2} = \mathcal{I}_{\{k=c\}}\frac{1}{\sigma^2}\left[\frac{\partial h_{ic}}{\partial \beta_k}(\gamma_i - \beta_k) - h_{ic}\right], \text{ and } \frac{\partial^2 h_{ic}}{\partial \beta_k \beta_{k'}} = 0 \text{ for } k \neq k'$$

$$\frac{\partial h_{ic}}{\partial \sigma} = w_{ic}\left[-\frac{1}{\sigma^2}e^{-\frac{(\gamma_i-\beta_c)^2}{2\sigma^2}} + \frac{1}{\sigma}e^{-\frac{(\gamma_i-\beta_c)^2}{2\sigma^2}}\frac{(\gamma_i-\beta_c)^2}{\sigma^3}\right] = -\frac{h_{ic}}{\sigma} + \frac{h_{ic}}{\sigma^3}(\gamma_i - \beta_c)^2$$

$$\frac{\partial^2 h_{ic}}{\partial \sigma^2} = -\left(\frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma - h_{ic}}{\sigma^2}\right) + (\gamma_i - \beta_c)^2\frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma^3 - 3h_{ic}\sigma^2}{\sigma^6}$$

$$\frac{\partial^2 h_{ic}}{\partial \beta_k \partial \sigma} = \left(\frac{\frac{\partial h_{ic}}{\partial \sigma}}{\sigma^2} - \frac{2h_{ic}}{\sigma^3}\right)(\gamma_i - \beta_s)s$$

Quantitative traits with covariates and shared variance

In this case, the h_{ic} function takes the form

$$h_{ic} = w_{ic}\frac{1}{\sigma}e^{-\frac{(\gamma_i-\psi_{is})^2}{2\sigma^2}}, \quad \text{where} \quad \psi_{is} = \beta_s + \sum_{p=1}^{P}\gamma_p z_{ip}.$$

Therefore,

$$\frac{\partial h_{ic}}{\partial \beta_k} = \mathcal{I}_{\{k=c\}}h_{ic}\frac{\gamma_i - \psi_{ic}}{\sigma^2}$$

$$\frac{\partial^2 h_{ic}}{\partial \beta_k^2} = I_{\{k=c\}}\frac{1}{\sigma^2}\left[\frac{\partial h_{ic}}{\partial \beta_k}(\gamma_i - \psi_{ic}) - h_{ic}\right], \text{ and } \frac{\partial^2 h_{ic}}{\partial \beta_k \beta_{k'}} = 0, \text{ for } k \neq k'$$

$$\frac{\partial h_{ic}}{\partial \sigma} = -\frac{h_{ic}}{\sigma} + \frac{h_{ic}}{\sigma^3}(\gamma_i - \psi_{ic})^2$$

$$\frac{\partial^2 h_{ic}}{\partial \sigma^2} = -\left(\frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma - h_{ic}}{\sigma^2}\right) + (\gamma_i - \psi_{ic})^2\frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma^3 - 3h_{ic}\sigma^2}{\sigma^6}$$

$$\frac{\partial^2 h_{ic}}{\partial \beta_k \partial \sigma} = I_{\{k=c\}}\left(\frac{\frac{\partial h_{ic}}{\partial \sigma}}{\sigma^2} - \frac{2h_{ic}}{\sigma^3}\right)(\gamma_i - \psi_{ic})$$

$$\frac{\partial^2 h_{ic}}{\partial \gamma_p \partial \sigma} = \left(\frac{\frac{\partial h_{ic}}{\partial \sigma}}{\sigma^2} - \frac{2h_{ic}}{\sigma^3}\right)(\gamma_i - \psi_{ic})z_{ip}$$

$$\frac{\partial^2 h_{ic}}{\partial \gamma_p \partial \beta_k} = I_{\{k=c\}}\frac{z_{ip}}{\sigma^2}\left(\frac{\partial h_{ic}}{\partial \beta_j}(\gamma_i - \psi_{ic}) - h_{ic}\right)$$

$$\frac{\partial h_{ic}}{\partial \gamma_p} = \frac{\partial h_{ic}}{\partial \sigma^2}(\gamma_i - \psi_{ic})z_{ip}$$

$$\frac{\partial^2 h_{ic}}{\partial \gamma_p^2} = \left(\frac{\partial h_{ic}}{\partial \gamma_p}\right)^2\frac{1}{h_{ic}} - h_{ic}\frac{z_{ip}^2}{\sigma^2}, \text{ and } \frac{\partial^2 h_{ic}}{\partial \gamma_p \gamma_{p'}} = \frac{\partial h_{ic}}{\partial \gamma_p}\frac{\partial h_{ic}}{\partial \gamma_{p'}}\frac{1}{h_{ic}} - h_{ic}\frac{z_{ip}z_{ip'}}{\sigma^2} \text{ for } p \neq p'$$

Trend test

In this situation we can write the linear predictor of equation (18) as

$$\psi_{ic} = \beta_1 + \beta_1(c-1).$$

In other words, β_1 plays the role of an intercept and β_2 is the slope. In this case, we consider that both β_1 and *beta*2 are shared for each latent class. In this situation, bearing in mind that $h_{ic} = w_{ic}\frac{e^{\gamma_i\psi ic}}{1+e^{\psi ic}}$, for the discrete traits, we have that

$$\frac{\partial h_{ic}}{\partial \beta_k} = h_{ic}x_{ikc}(\gamma_i - p_{ic}), \tag{19}$$

and

$$\frac{\partial^2 h_{ic}}{\partial \beta_k \partial \beta_{k'}} = x_{ikc}\frac{\partial h_{ic}}{\partial \beta_{k'}}(\gamma_i - p_{ic}) - x_{ikc}x_{ik'c}h_{ic}(p_{ic} - p_{ic}^2). \tag{20}$$

For quantitative traits, where $h_{ic} = w_{ic}\frac{1}{\sigma}e^{-\frac{(\gamma_i-\psi_{ic})^2}{2\sigma^2}}$, we have that

$$\frac{\partial h_{ic}}{\partial \beta_k} = h_{ic} x_{ikc} \frac{\gamma_i - \psi_{ic}}{\sigma^2}, \quad (21)$$

and

$$\frac{\partial^2 h_{ic}}{\partial \beta_k \partial \beta_{k'}} = x_{ikc} \frac{\partial h_{ic}}{\partial \beta_{k'}} \frac{\gamma_i - \psi_{ic}}{\sigma^2} - x_{ikc} x_{ik'c} \frac{h_{ic}}{\sigma^2}. \quad (22)$$

For the variance, we have that

$$\frac{\partial h_{ic}}{\partial \sigma} = -\frac{h_{ic}}{\sigma} + \frac{h_{ic}}{\sigma^3}(\gamma_i - \psi_{ic})^2, \quad (23)$$

$$\frac{\partial^2 h_{ic}}{\partial \sigma^2} = -\left(\frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma - h_{ic}}{\sigma^2} \right) + (\gamma_i - \psi_{ic})^2 \frac{\frac{\partial h_{ic}}{\partial \sigma}\sigma^3 - 3h_{ic}\sigma^2}{\sigma^2}, \quad (24)$$

and

$$\frac{\partial^2 h_{ic}}{\partial \beta_k \partial \sigma} = x_{ikc} \left(\frac{\frac{\partial h_{ic}}{\partial \sigma}}{\sigma^2} - \frac{2h_{ic}}{\sigma^3} \right) (\gamma_i - \psi_{ic}). \quad (25)$$

Additional material

Additional file 1
Tables and figures for more scenarios of simulation studies.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-172-S1.pdf]

Acknowledgements

The first author would like to thank Xavier Bassagaña for his comments and helpful conversations about the model proposed. Gavin Lucas is also acknowledged for his comments on a last version of the manuscript. The authors also want to thank helpful comments on how to analyze aCGH data given by one of the reviewers. This work was supported by the Spanish Ministry for Science and Innovation [MTM2008-02457 to JRG and SAF2008-00357 to XE]; and the European Commission [AnEUploidy project; FP6-2005-LifeSciHealth contract #037627].

References

1. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM and Eichler EE: **Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.** *Am J Hum Genet* 2006, **79(2):**275–290.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Grata-cos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Es-tivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW and Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444(7118):** 444–454.
3. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE and Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80:**91–104.
4. Feuk L, Carson AR and Scherer SW: **Structural variation in the human genome.** *Nat Rev Genet* 2006, **7(2):**85–97.
5. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME and Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315(5813):**848–853.
6. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ and Ahuja SK: **The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.** *Science* 2005, **307(5714):** 1434–440.
7. Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T and Campion D: **APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy.** *Nat Genet* 2006, **38:**24–26.
8. Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P and Ferec C: **Hereditary pancreatitis caused by triplication of the trypsinogen locus.** *Nat Genet* 2006, **38(12):**1372–1374.
9. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F and G P: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.** *Nucleic Acids Res* 2002, **30(12):**e57.
10. González J, Carrasco J, Armengol L, Villatoro S, Jover L, Yasui Y and Estivill X: **Probe-specific mixed-model approach to detect copy number differences using multiplex ligation-dependent probe amplification (MLPA).** *BMC Bioinformatics* 2008, **9:**261.
11. Engert S, Wappenschmidt B, Betz B, Kast K, Kutsche M, Hellebrand H, Goecke T, Kiechle M, Niederacher D, Schmutzler R and Meindl A: **MLPA screening in the BRCA1 gene from 1,506 German hereditary breast cancer cases: novel deletions, frequent involvement of exon 17, and occurrence in single early-onset cases.** *Hum Genet* 2008, **29(7):**948–958.
12. Hansen T, Jonson L, Albrechtsen A, Andersen M, Ejlertsen B and Nielsen F: **Large BRCA1 and BRCA2 genomic rearrangements in Danish high risk breast-ovarian cancer families.** *Breast Cancer Res Treat* 2008 in press.
13. Aitman T, Dong R, Vyse T, Norsworthy P, Johnson M, Smith J, Mangion J, Roberton-Lowe C, Marshall A, Petretto M, Hodges E, Bhangal G, Patel S, Sheehan-Rooney K, Duda M, Cook P, Evans D, Domin J, Flint J, Boyle J, Pusey C and Cook H: **Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans.** *Nature* 2006, **439(7078):**851–855.
14. Fellermann K, Stange D, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins C, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B and Stange E: **A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon.** *Am J Hum Genet* 2006, **79(3):** 439–48.
15. Ionita-Laza I, Rogers AJ, Lange C, Raby BA and Lee C: **Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis.** *Genomics* 2009, **93:**22–26.
16. Fraley C and Raftery AE: **How many clusters? Which clustering method? Answers via model-based cluster analysis.** *The Computer Journal* 1998, **41:**578–588.
17. Picard F, Robin S, Lebarbier E and Daudin JJ: **A segmentation/clustering model for the analysis of array CGH data.** *Biometrics* 2007, **63(3):**758–766.
18. Wiel van de MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM and Ylstra B: **CGHcall: calling aberrations for array CGH tumor profiles.** *Bioinformatics* 2007, **23(7):**892–894.
19. Leisch F: **A general framework for finite mixture models and latent class regression in R.** *Journal of Statistical Software* 2004, **11(8):** 1–18.
20. Du J: **Combined Algorithms for Fitting Finite Mixture Distributions.** *PhD thesis* McMaster University, Ontario, Canada; 2002.
21. Bashir S and Duffy S: **The correction of risk estimates for measuremente error.** *Ann Epidem* 1993, **7:**156–164.

22. Davidov O, Faraggi D and Reiser B: **Misclassification in logistic regression with discrete covariates.** *Biometrical Journal* 2003, **5:**541–553.
23. Greenland S: **Basic methods for sensitivity analysis of biases.** *Int J Epi* 1996, **25:**1107–1115.
24. Spiegelman D, Rosner B and Logan R: **Estimation and inference for logistic regression with covariate missclassification and measurement error, in main study/validation study designs.** *J Am Stat Assoc* 2000, **95:**51–61.
25. **CREAL's web-page.** http://www.creal.cat/jrgonzalez/software.htm.
26. Wiel van de M and van Wieringen W: **CGHregions: dimension reduction for array CGH data with minimal information loss.** *Cancer Informatics* 2007, **2:**55–63.
27. Benjamini Y and Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Roy Statist Soc Ser B* 1995, **57:**289–300.
28. Sarkar S: **False discovery and false nondiscovery rates in single-step multiple testing procedures.** *The Annals of Statistics* 2006, **34:**394–415.
29. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A and Gray JW: **A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.** *Cancer Cell* 2006, **10(6):**515–527.
30. **Bioconductor's web-page.** http://www.bioconductor.org/.
31. M Neve et al in Gray Lab at LBL: *Neve2006: expression and CGH data on breast cancer cell lines. [R package version 0.1.6].*
32. van Wieringen WN and Wiel van de MA: **Nonparametric testing for DNA copy number induced differential mRNA gene expression.** *Biometrics* 2009, **65:**19–29.

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Genetic Association Analysis and Meta-Analysis of Imputed SNPs in Longitudinal Studies

Isaac Subirana[1,2,3] and Juan R González[1,4,5]*

[1]CIBER Epidemiology and Public Health (CIBERESP), Spain; [2]Cardiovascular Epidemiology & Genetics Group, Inflammatory and Cardiovascular Disease Programme, IMIM, Parc de Salut Mar, Spain; [3]Department of Statistics, University of Barcelona, Barcelona, Spain; [4]Center for Research in Environmental Epidemiology (CREAL), Spain; [5]Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

Received 31 August 2012; Revised 10 January 2013; accepted revised manuscript 5 February 2013.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21719

ABSTRACT: **In this paper we propose a new method to analyze time-to-event data in longitudinal genetic studies. This method address the fundamental problem of incorporating uncertainty when analyzing survival data and imputed single-nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS). Our method incorporates uncertainty in the likelihood function, the opposite of existing methods that incorporate the uncertainty in the design matrix. Through simulation studies and real data analyses, we show that our proposed method is unbiased and provides powerful results. We also show how combining results from different GWAS (meta-analysis) may lead to wrong results when effects are not estimated using our approach. The model is implemented in an R package that is designed to analyze uncertainty not only arising from imputed SNPs, but also from copy number variants.**
Genet Epidemiol 00:1–13, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: **imputed SNP; longitudinal studies; GWAS; genetic association**

Introduction

Genome-wide association studies (GWAS) have found a large number of single-nucleotide polymorphisms (SNPs) associated with several complex diseases such as Alzheimer [Rovelet-Lecrux et al., 2006], cancer [Xing et al., 2011], or cardiovascular diseases [Myocardial Infarction Genetics Consortium, 2009], among others. Some of these GWAS are based on cohort studies such as Framingham [Granada et al., 2011], INMA [Morales et al., 2011], and CHARGE [Fornage et al., 2011]. Therefore, when longitudinal data is available researchers are interested not only in discovering genes related to susceptibility but also in determining genes related to progression [Frey et al., 2006b; González et al., 2011], survival [Frey et al., 2006a; Heist et al., 2007], time to relapse or recurrence [Wojnar et al., 2009; Xing et al., 2011], or response to treatment [Alakus et al., 2009; Ueno et al., 2008]. Conversely to case-control studies, some of these cohort studies have found genetic variants with very high effects. For example, Frey et al. [2005] discovered that patients diagnosed with colorectal cancer who carried CC genotype of T393C polymorphism were at highest risk for death (hazard ratio [HR], 12.1) compared with TT genotypes. On the other hand, González et al. [2011] estimated that the probability of multiple sclerosis progression for those individuals who carried two copies of the CD24 gene was 3.4 times higher than for those who carried the frequent allele. Probably, these high HR are overestimated and have to be replicated. Nonetheless, these HR are normally larger than the odds ratio (OR) found in case-control studies that lay between 1.1 and 1.5.

Statistical methods to analyze time-to-event data in genetic studies are based on standard survival analysis techniques. The hazard risk of observing the event of interest for a given SNP is normally estimated using either a parametric (Weibull, log-normal, ...) [Del Greco et al., 2011; Park et al., 2010] or a semiparametric model (Cox regression) [Morgan et al., 2011; Takeuchi et al., 2009]. To improve power, current GWAS combine information across cohorts. Each study may use different platforms to get genotype information, resulting in different numbers of genotyped SNPs. Therefore, in order to analyze the same number of SNPs for all studies, imputed data are normally obtained from each study. These SNPs are normally estimated with uncertainty; hence, standard statistical techniques cannot be used to assess association between the SNPs and the outcome. Imputing methods provide for each SNP a vector with three probabilities, corresponding to each genotype, that indicates the probability of having 0, 1, or 2 copies of the rare allele. The simplest method to estimate the HR for an imputed SNP is to consider the genotype with the highest probability as if it were observed without uncertainty (best guess). Aulchenko et al. [2010] pointed out that this approach, which does not incorporate the uncertainty in the model, is biased and underpowered.

Statistical models to analyze survival data with imputed SNPs should incorporate uncertainty. To our knowledge, the only existing method to do so in cohort studies is the one

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Dr. Juan R. González, Center for Research in Environmental Epidemiology (CREAL), Room 188, Barcelona Biomedical Research Park (PRBB), Plaza Charles Darwin s/n, Barcelona 08003, Spain. E-mail: jrgonzalez@creal.cat

proposed by Aulchenko et al. [2010], which incorporates the probabilities of each genotype of the imputed SNP in the design matrix. SNPTEST software [Marchini et al., 2007] also implements methods to incorporate uncertainty, but only for case-control studies and does not accommodate time-to-event responses. Parametric or semiparametric methods can then be used to estimate the HR. The authors argue that this procedure gives unbiased results. However, they did not provide any justification, maybe due to the fact that this method is being widely used in most of the recent papers that use logistic regression models [Kooner et al., 2011; Trompet et al., 2011]. Theoretically, this approach should work in linear models (because it is not biased) but its use cannot be correct when logistic regression is used for case-control studies or with any other nonlinear regression, as in survival data models. In order to overcome this difficulty, we propose an alternative approach based on a model that incorporates uncertainty in the likelihood function. The main aim of this paper is to investigate whether a best-guess approach, to incorporate the uncertainty as covariates DOSAGE, and our proposed method can be used to analyze follow-up cohort genetic studies when analyzing imputed SNPs. We are also interested in assessing the impact of the proposed method (mainly in terms of bias of the pooled HR) when performing meta-analysis by combining HR estimated for imputed SNPs in different studies.

The remainder of the paper is organized as follows: First section, "Methods," briefly describes the main existing programs to get imputed SNPs. We also provide the likelihood function for the three studied methods: best guess, uncertainty as covariates, and our new proposed method. The second section, "Simulation Study," describes the simulation studies carried out to check whether each method is biased when estimating HRs and to assess power under different scenarios (allele frequency, effect, and uncertainty degree). Also we evaluate the impact of combining HR of imputed SNPs in meta-analyses. Simulations are based on imputed SNPs from the Framingham Study cohort (http://www.framingham.com/heart/) to mimic real situations. The third section, "Application to Real Data," presents the results from a real data set using imputed SNPs as well as phenotypes from the Framingham Study cohort. Finally, we present our conclusions.

Methods

Imputation programs

Programs to impute SNPs include PLINK, IMPUTE, MACH, and fast-PHASE. They have been used in different studies, mainly in case-control settings [Meyre et al., 2009; Nair et al., 2009]. Each program uses its own criteria, data elements, and algorithms to perform the imputation of nongenotyped SNPs. A comprehensive comparison of all these methods can be found in Biernacka et al. [2009]. All of them provide a quality score that gives an idea about how well each genotype is imputed. Most of them return a vector with three probabilities for each individual and for each imputed SNP, i.e., the probability that a particular individual has the genotype "aa," "aA," or "AA" (0, 1, or 2 risk alleles, respectively). There also exist other imputation methods that only return the most likely genotype for each individual. This result is less informative than providing the three probabilities because uncertainty is not taken into account. It is obvious that, for instance, having a 99% probability of presenting "AA" genotype is not the same risk as having a 51% probability. In both situations "AA" will be considered as the most likely genotype but in the second case we are not so sure that it is true, i.e., there is much more uncertainty when assigning the "AA" genotype. Therefore, we strongly recommend discarding methods that do not return probabilities because uncertainty is important when assessing association [Aulchenko et al., 2010].

Model to Analyze Imputed SNPs in Longitudinal Studies

We consider that each individual is being followed up over a study period $[0, C]$, where C may represent time of study termination or some other right-censoring variable (e.g., another event different from the one under study). The response variable is the time until the event of interest, T, that cannot be observed if the event appears after the censoring time. Therefore, the observable response variable is (Y, δ), where

$$Y = \min(T, C),$$

and

$$\delta = \begin{cases} 1 & \text{if} \quad Y \geq \mathrm{T} \\ 0 & \text{if} \quad Y < T. \end{cases} \tag{1}$$

The probability of having $k = 0, 1, 2$ copies of the risk allele for a given imputed SNP can be denoted by

$$\mathrm{P}(\mathrm{SNP} = k) \equiv w_k.$$

Other possible nongenetic covariates (in GWAS we need at least to consider principal components), $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ can also be observed. Consequently, if the study has n individuals we will observe the following data

$$\boldsymbol{D} \equiv \{(Y_i, \delta_i), (w_{i0}, w_{i1}, w_{i2}), \boldsymbol{Z}_i, i = 1, 2, \ldots, n\}, \tag{2}$$

where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})$ is the vector of nongenetic covariates observed for the ith individual. We propose to model the hazard rate process at time t conditionally on the imputed SNP and the observed covariates via the following Weibull model:

$$\lambda(t \mid w_{ik} Z) = \phi t^{\phi-1} \exp(\alpha + \beta k + \boldsymbol{Z}\boldsymbol{\gamma}'), \tag{3}$$

where ϕ denotes the Weibull shape parameter. Then, using the total probability theorem, it is straightforward to see that the likelihood function for the observable data is:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \sum_{k=0}^{2} \mathrm{P}(Y_i, \delta_i \mid \mathrm{SNP}_i = k, \boldsymbol{Z}_i; \boldsymbol{\Theta}) w_{ik}, \tag{4}$$

where $\mathrm{P}(Y_i, \delta_i \mid \cdot)$ is the density function, $f(Y_i \mid \cdot)$, if $\delta_i = 1$ or the survival function, $S(Y_i \mid \cdot)$, if $\delta_i = 0$ and

$\boldsymbol{\Theta} = (\alpha, \beta, \gamma_1, \ldots, \gamma_p, \phi)$ is the vector of parameters, α corresponds to the constant in the linear predictor, β denotes the coefficient (e.g., log-HR) for the imputed SNP assuming an additive effect (other genetic models can also be considered), $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ encode the effects of covariates, and ϕ is the shape parameter.

By maximizing equation (4), we ensure that model parameter estimates are asymptotically nonbiased and efficient by maximizing the likelihood function. Note that our formulation considers that w_{ik} are known. In particular, we propose to obtain them from any of the existing imputed methods. This approach was also adopted in the context of case-control studies, with excellent results [Gonzalez et al., 2009]. The Newton-Rapshon iterative algorithm can be used to compute the maximum likelihood parameter estimates, $\hat{\boldsymbol{\Theta}}$.

Model Parameter Estimation and Likelihoods

Two existing approaches can be used to maximize (4). Both of them depend on how uncertainty is incorporated into the model. These two methods have some drawbacks mainly due to the fact that they do not maximize the true likelihood function. To overcome this problem, we propose a third approach to maximize the observed likelihood. The three methods can be summarized as follows:

1. Best guess approach (`Naive`): this classical method considers the most probable genotype for each individual and proceeds as if it is the real genotype. This method can be used even if the imputation method only provides the most likely genotype. This approach corresponds to the classical survival models used when analyzing observed SNPs.
2. Probabilities as covariates (`DOSAGE`): this method addresses uncertainty by introducing it as a covariate in the design matrix. If an additive inheritance model is assumed, a linear transformation of the three probabilities results in a single value per individual, ranging from 0 to 2. This quantity is known as *dosage*.
3. Latent Class (`LC`): this method incorporates the uncertainty in the likelihood function (4). Model parameters can be estimated using the Newton-Raphson algorithm using score and Hessian matrices (derived in the present paper—see Appendix). The `LC` approach has been incorporated to the `R` software [R Development Core Team, 2012] package `CNVassoc` [Subirana et al., 2011].

Next, we introduce the likelihood function used to obtain model parameters for the `Naive`, `DOSAGE`, and `LC` methods and discuss the adequacy of each one.

Best Guess Approach (`Naive`)

This approach does not maximize (4). Rather, it maximizes the function that results from replacing the probabilities w_{ik} by 0 and 1 depending on the maximum probability observed for each genotype. That is, $w_{ik} = 1$ if w_{ik} is equal to $\max_k(w_{ik})$ and 0 otherwise. Therefore, this method maximizes this *pseudo-likelihood* function:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \mathrm{P}(Y_i, \delta_i \mid \mathrm{SNP}_i = k_i^*, \boldsymbol{Z}_i; \boldsymbol{\Theta}), \qquad (5)$$

where k_i^* is the most likely genotype for the ith individual.

It is clear that (5) will be similar to (4) when $\max_k(w_{ik})$ is close to 1 for all individuals (i.e., low uncertainty). However, when uncertainty is moderate or high, (5) will be very different from (4), leading to biased parameter estimates.

Probabilities as Covariates (`DOSAGE`)

`DOSAGE` approach considers the probabilities w_{ik} as covariates by incorporating them into the design matrix. In this case, it maximizes this *pseudo-likelihood* function:

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \mathrm{P}(Y_i, \delta_i \mid d_i, \boldsymbol{Z}_i; \boldsymbol{\Theta}), \qquad (6)$$

where d_i is a covariate (or covariates) that depends on the inheritance model as follows:

$$d_i \equiv \begin{cases} 0 * w_{i0} + 1 * w_{i1} + 2 * w_{i2} & \text{for additive models (\textit{dosage})} \\ 0 * w_{i0} + 0 * w_{i1} + 1 * w_{i2} & \text{for recessive models} \\ 0 * w_{i0} + 1 * w_{i1} + 1 * w_{i2} & \text{for dominant models} \\ (d_{i1}, d_{i2}) = (w_{i1}, w_{i2}) & \text{for model free} \end{cases}.$$

Note that, for the model free, two dummy variables are needed, which in this case are the probability of having 1 copy and the probability of having 2 copies, taking 0 copies as the reference category group. The coefficient of d_i can be interpreted as the effect of the SNP. For the model free, the coefficient of d_{i1} and d_{i2} are the effect of having 1 copy and 2 copies vs. 0 copies, respectively.

The quantity d_i is also known as *dosage* effect when an additive inheritance model is assumed. This quantity is a standard output for most impute software (IMPUTE, PLINK, etc.). It is important to note that although it is easy to compute the *dosage* from the probabilities w_{ik}, it is not possible to obtain the probabilities from the *dosage*.

Latent Class (`LC`)

The two previous methods have some drawbacks because they do not maximize the proper likelihood (4). Therefore, nonbiased and efficient estimates are not guaranteed. Nonetheless, equation (4) can be maximized using an iterative procedure and the analytic score and Hessian function that are derived in the Appendix.

Although `Naive` and `DOSAGE` approaches maximize their corresponding likelihood functions using standard algorithms, `LC` requires a more sophisticated one because (4) is a product of a sum that is not simplified when taking the logarithms, as in (5) or (6).

It is important to note that, although `Naive` and `DOSAGE` strategies rely on incorporating the imputed SNP

probabilities as covariates and fit an standard regression model including a semiparametric Cox regression model, LC cannot only support a fully parametric model such as a Weibull distributed response. However, Weibull distribution is flexible enough to fit on most of real data.

Simulation Study

We carried out simulation studies to examine properties of estimators of parameters using Naive, DOSAGE, and LC approaches. Our primary aim was to ascertain the robustness of hazard rate estimators of imputed SNPs, $\beta = \log(\text{HR})$. Our second aim was to assess the impact when combining HR estimated in different GWAS using imputed SNPs. We illustrate both simulations in the next sections.

Accuracy to Estimate HR

In order to mimic available imputed SNP data, 5,000 data sets were simulated using the following scheme:

1. **Imputed SNPs:** Imputed SNPs were randomly selected from the Framingham Heart Study cohort (http://www.framingham.com/heart/), a sample with more than 8,477 individuals containing about a million genotyped SNPs and approximately 2.5 million imputed SNPs. Different minor allele frequencies (MAF) and different degrees of imputation uncertainty were considered. The uncertainty measure used was R^2 defined as the correlation between actual genotype and most probable genotype. This definition of R^2 is incorporated by imputation software MACH [Li et al., 2010], which was used to impute SNPs in the Framingham cohort. R^2 ranges from 0 (no information, huge uncertainty) to 1 (complete information, no uncertainty). We selected 100 SNPs from each combination of MAF in [0.01, 0.05), [0.05, 0.15), [0.25, 0.35), and [0.45, 0.50) and R^2 in [0.05, 0.15), [0.25, 0.35), [0.45, 0.55), [0.65, 0.75), and [0.85, 0.95). Table 1 shows the total number of imputed SNPs from the Framingham cohort in each of these combination of R^2 by MAF bins.
2. **Nonobserved genotypes:** Because imputed Framingham genotype data do not include actual genotype, they have been simulated as follows: from each individual and each imputed SNP, one of the three possible genotypes were sampled from the imputed probabilities vector. Therefore, we proceeded differently from Zheng et al. [2011]. Of course, our strategy in obtaining the actual genotype assumes that the imputed probabilities are correct. We considered two reasons for adopting this strategy: (1) it saves a lot of time in avoiding the need to perform SNP imputation again (Framingham data contains already imputd SNPs); and (2) it removes the possible effect of imputation bias in assessing the association test, which is not the aim of our study.
3. **Response (time-to-event) simulation:** Two different strategies have been adopted when simulating the response: (1) assuming a Weibull distribution; or (2) taking the empirical observed time-to-coronary event from Framingham cohort. In the two scenarios, baseline distribution functions has been obtained for the three genotypes assuming a proportional hazard:

$$F(t) = 1 - S(t)^{\exp(\beta \text{SNP})},$$

where F is the distribution function, S is the baseline survival function, and SNP is the genotype coded as (0, 1, 2) which represents the number of risk alleles and β is the logarithm of HR.
For (1), scale and shape parameters for S has been obtained from the estimates using the Framingham data, whereas for (2) S has been obtained as the Kaplan-Meier estimates also from Framingham data. Different scenarios for the SNP effect (i.e., HR) has been performed, ranging from low to very high (1.5, 2, 2.5, 3, 3.5, and 4), also including an scenario with no effect (HR = 1).
For the first strategy (a Weibull distributed response), the following strategy has been followed to simulate censored events: simulated values exceeding the maximum follow-up time observed in Framingham have been censored; in addition, some extra values within the follow-up interval has been censored, too, in order to achieve a similar number of events as observed in the Framingham cohort. In (2), two Kaplan-Meier curves have been performed, one taking the observed events and the other taking the censored events. From the first one, time-to-event values, T, have been generated, and censoring times, C, from the second one. The observed times, Y, have been defined as $Y = \min\{C, T\}$. Finally, if $C > T$, the value has been considered censored, and noncensored otherwise (i.e., $C \leq T$). For all strategies, the rate of noncensored events has been similar to the observed coronary events rate in the Framingham cohort (i.e., 8% approximately).
Results from empirical response are shown in the Supplementary Materials.
4. **Estimation:** For each simulated data element (time-to-response, nonobserved genotype, and imputed SNP probabilities), four models were estimated: a proportional hazard Cox regression model using the actual (nonobserved) genotype (True), the Naive model using the most probable SNP number of alleles, (DOSAGE) with the *dosage* formulation as a predictor, and LC using imputed SNP probabilities. Finally, an additive mode of

Table 1. Distribution of imputed SNPs according to combination of R^2 and MAF from the Framingham cohort

	MAF (%)			
R^2 (%)	1–5	5–15	25–35	45–50
5–15	9,322	3,862	1,513	600
25–35	13,279	9,708	3,893	1,537
45–55	13,258	14,892	6,797	2,795
65–75	17,687	26,508	14,362	5,773
85–95	56,875	93,339	68,742	29,429

inheritance was assumed for all the fitted models. The `True` model was used as the gold standard and can be used as a reference. Note that, it is not feasible to use the proportional hazard Cox model for `LC` approach. Therefore, the first three approaches uses a semiparametric model (Cox regression) although `LC` uses a full parametric model (Weibull regression).

In each simulation and model, the following measures were computed: (1) Bias: defined as the difference of the expected and the true HR; (2) MSE: mean-squared error, defined as the expected squared difference between estimated log-HR and the true log-HR; (3) Power: power of detecting an associated SNP; (4) Coverage: probability that the 95% confidence interval includes the true HR.

Impact when Combining Hazard Ratios for GWAS Meta-Analysis

Using simulated data from noncensored Weibull distribution, five data sets are taken as if they correspond to five different GWAS for which the HR was estimated using the true SNP. This will be considered as the gold-standard results. Another five data sets were used to compute the HR using the three compared approaches (`Naive`, `DOSAGE`, and `LC`). Pooled HR was fitted using the `rmeta` package assuming a random-effect model. We repeated this process 5,000 times. For each simulation, pooled HR, heterogeneity *P*-value and τ^2 estimation were recorded. The heterogeneity *P*-value tests whether there is an excess of heterogeneity, although τ^2 is a measure of that heterogeneity.

From the same simulated data, we have taken 10 data sets at random, as if they correspond to 10 different GWAS estimated using each of the three approaches as well as the true SNP. For each of the three approaches and the true SNP, a meta-analysis has been computed to assess the possible bias in estimating the pooled HR.

Results of Simulation Study

We computed the bias, power, MSE, and coverage after estimating HR for the imputed SNP. All the results have been plotted in different figures including different panels corresponding to the different scenarios when varying MAF and R^2 index. The figure in each panel shows the results for different HRs. In the discussion of the simulation results, we first focus on the estimation of HR. Then, we analyze the results to determine the effect of each of the three methods when computing pooled HR in meta-analysis settings.

We are illustrating the results for censored Weibull response. For this case, we computed the bias (Figure 1), power (Figure 2), MSE (Supplementary Fig. S1), and coverage (Supplementary Fig. S2) after estimating HR for the imputed SNP. All these figures include different panels corresponding to the different scenarios when varying MAF and R^2 index. The figure in each panel shows the results for different HRs.

Figure 1 shows that the estimated HR using the `LC` approach is almost identical to those estimated using the gold-standard method (true SNP) in all scenarios. The `DOSAGE` method is biased for those HRs larger than 2.5. For this model, we can also observe that the more uncertainty and the higher the MAF, the more bias exists. For instance, when the R^2 index is 0.3 and MAF is equal to 0.5, `DOSAGE` reports biased results for HR larger than 2. Finally, as one may expect, the `Naive` approach is completely biased for all scenarios.

Figure 2 shows the power of each approach for a GWAS significance level assuming 2.5 million of imputed SNPs ($\alpha = 2 \cdot 10e-8$). We can observe that `LC` and `DOSAGE` have similar power to detect association between the imputed SNP and the event of interest, although `Naive` method is less powerful than the other two when uncertainty increases.

Regarding to MSE, Supplementary Fig. S1 indicates the `LC` approach as the best one, although `DOSAGE` achieve almost the same accuracy. `Naive` behaves much worse than than `LC` and `DOSAGE` methods in almost all scenarios except when uncertainty is low ($R^2 > 85\%$).

Finally, `LC` also is the best approach with regard to coverage. Supplementary Fig. S2 shows that our proposed method has an observed coverage about 95% in all scenarios and for all HRs although the coverage dramatically decreases when HR increases for `DOSAGE` and `Naive` approaches.

Very similar results are obtained when simulating empirical distributed response in terms bias (Supplementary Fig. S3), power (Supplementary Fig. S4), accuracy (Supplementary Fig. S5), and coverage (Supplementary Fig. S6). It it remarkable to note that `LC` gives no bias estimates even when using a fully parametric Weibull model (Supplementary Fig. S3), and behaves as well as for censored Weibull-distributed response. Maybe this is due to empirical data could be fitted satisfactorily to a Weibull distribution.

Moving to meta-analysis simulations, Figures 3 and 4 summarize our main findings. Previous simulation studies have mainly shown that `Naive` and `DOSAGE` approaches are biased when estimating the risk of observing the event of interest. This has two main limitations when combining HR from different studies that use these approaches: first, the pooled HR is also biased and second, the variance of the pooled HR is inflated because the heterogeneity between studies increases and a random-effect model has to be used to pool the individual HRs. The first problem can be observed in Figure 3, which shows the simulation results when the true HR is equal to 2.5. Each panel corresponds to the meta-analysis results when combining the individual HRs estimated using each method. We can observe that the pooled HR is biased when using `Naive` and `DOSAGE`, although `LC` is not. The second problem is demonstrated in Figure 4, which shows the false positive rate of declaring that heterogeneity exists when combining five HRs estimated from one of the three approaches and five more HRs estimated from the `True` model. We simulated individual studies using the same expected HR. Therefore, the null hypothesis of heterogeneity between studies should not be rejected (e.g., false positive rate should be around 5%). However, Figure 4 shows that the type I error when using `Naive` and `DOSAGE` is inflated, and this inflation increases when HR is bigger. In practical terms, when heterogeneity is observed, a random-effects model has to be used to estimate

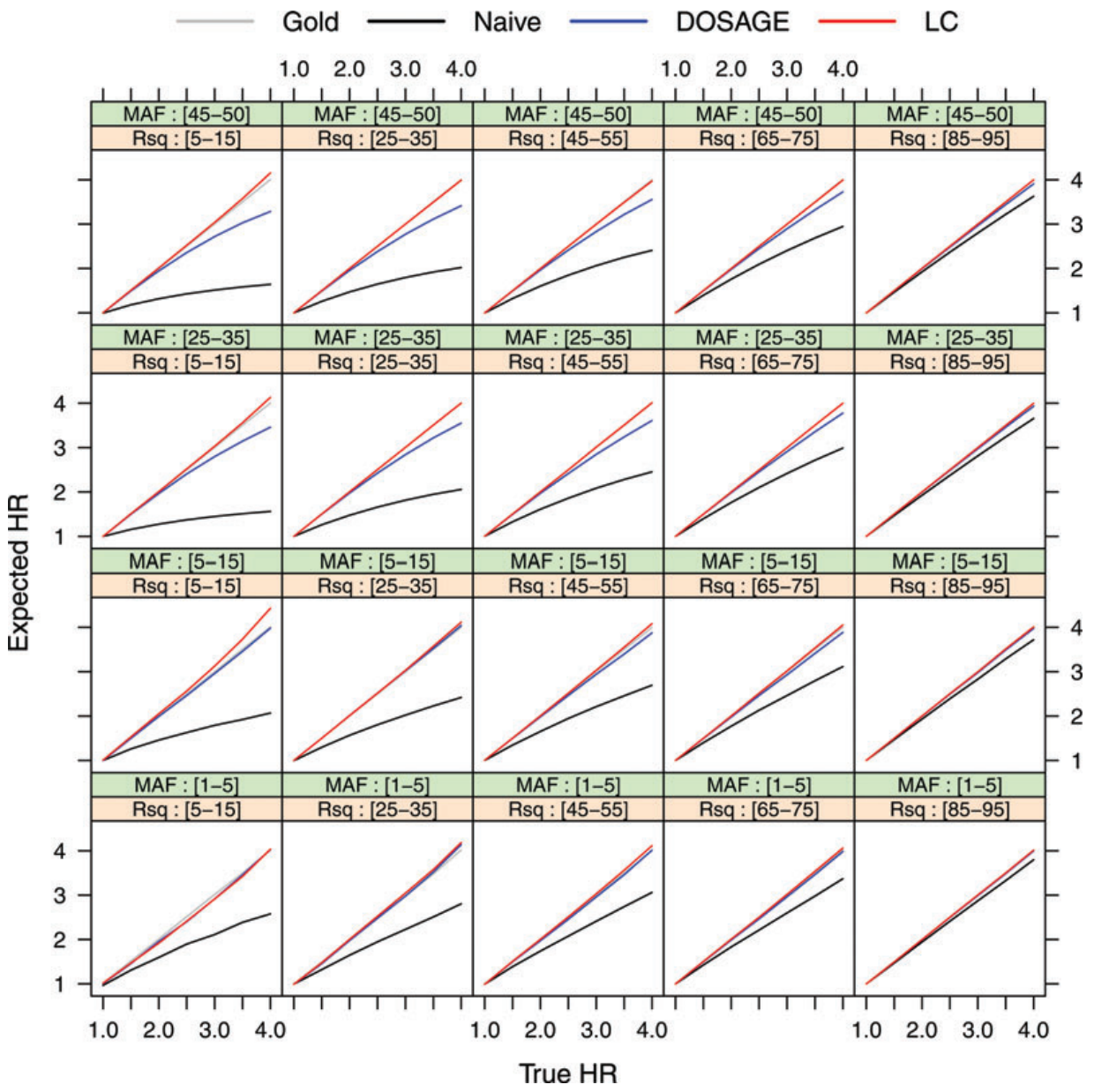


Figure 1. Hazard ratio according to minor allele frequency and uncertainty (R^2).

the pooled HR. This may increase the variance of pooled HR, decreasing the power to detect a significant pooled HR.

Finally, it must be emphasized that the failure rate (i.e., the number of iterations in which models have not converged) has been very low or null (data not shown). Only in very extreme scenarios (MAF below 5% and R^2 below 15%) has the failure rate been high (up to about 10% and 20% for `DOSAGE` and `LC`, respectively), and up to about 45% for `Naive` method. Therefore, in this scenario, the obtained results from the simulation study and meta-analysis study may not be reliable.

Application to Real Data

We illustrate our proposed method by analyzing SNPs and nongenetic variables from the Framingham Study data. We obtained access to phenotype and genotype (imputed SNPs) data under the Framingham Share initiative via the Database of Genotypes and Phenotypes (dbGaP, ncbi.nlm.nih.gov/dbgap; Project number 1534). From all imputed SNPs existing in the database, we discarded all the rare ones (with a MAF $\leq$ 1%) and the ones imputed with almost no uncertainty ($R^2 \geq$ 99%). Also, those SNPs with a very

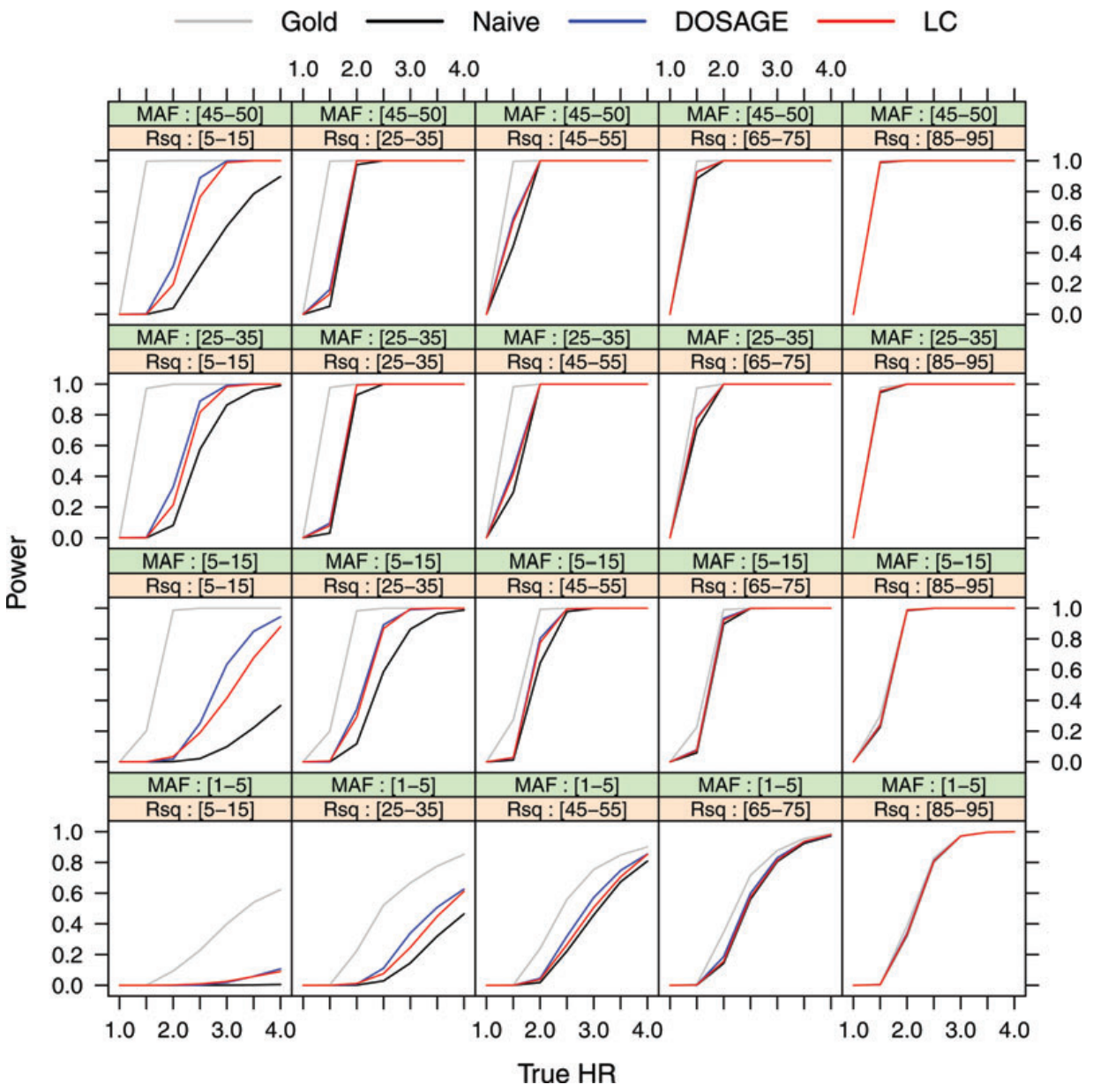


Figure 2. Power according to minor allele frequency and uncertainty (R^2), taking GWA significance level, α = 2·10e-8.

poor imputation quality ($R^2 \leq 10\%$) were removed. Finally, a set of 1,464,340 imputed SNPs were analyzed.

For each imputed SNP, the three approaches (`Naive`, `DOSAGE`, and `LC`) were fitted taking the time to coronary heart disease as the response. For the `Naive` and `DOSAGE` approaches, a proportional Cox regression was adjusted. As covariates, we considered age, sex, and the five first principal components to take into account family structure data. The first five principal components capture more than 80% of the variability. Finally, we analyzed 3,008 individuals from the Framingham offspring cohort with response information and all covariate data available.

To address an excess of familial or ethnic data structure of the data (not captured by the first five principal components), *P*-values were corrected by the genomic inflation factor λ (1.062, 1.069, and 1.058 for `Naive`, `DOSAGE`, and `LC` approaches, respectively). Note that, genomic inflation factor is very close to one. One explanation could be that the incorporation of the five principal components captures well the familiar structure in all three approaches.

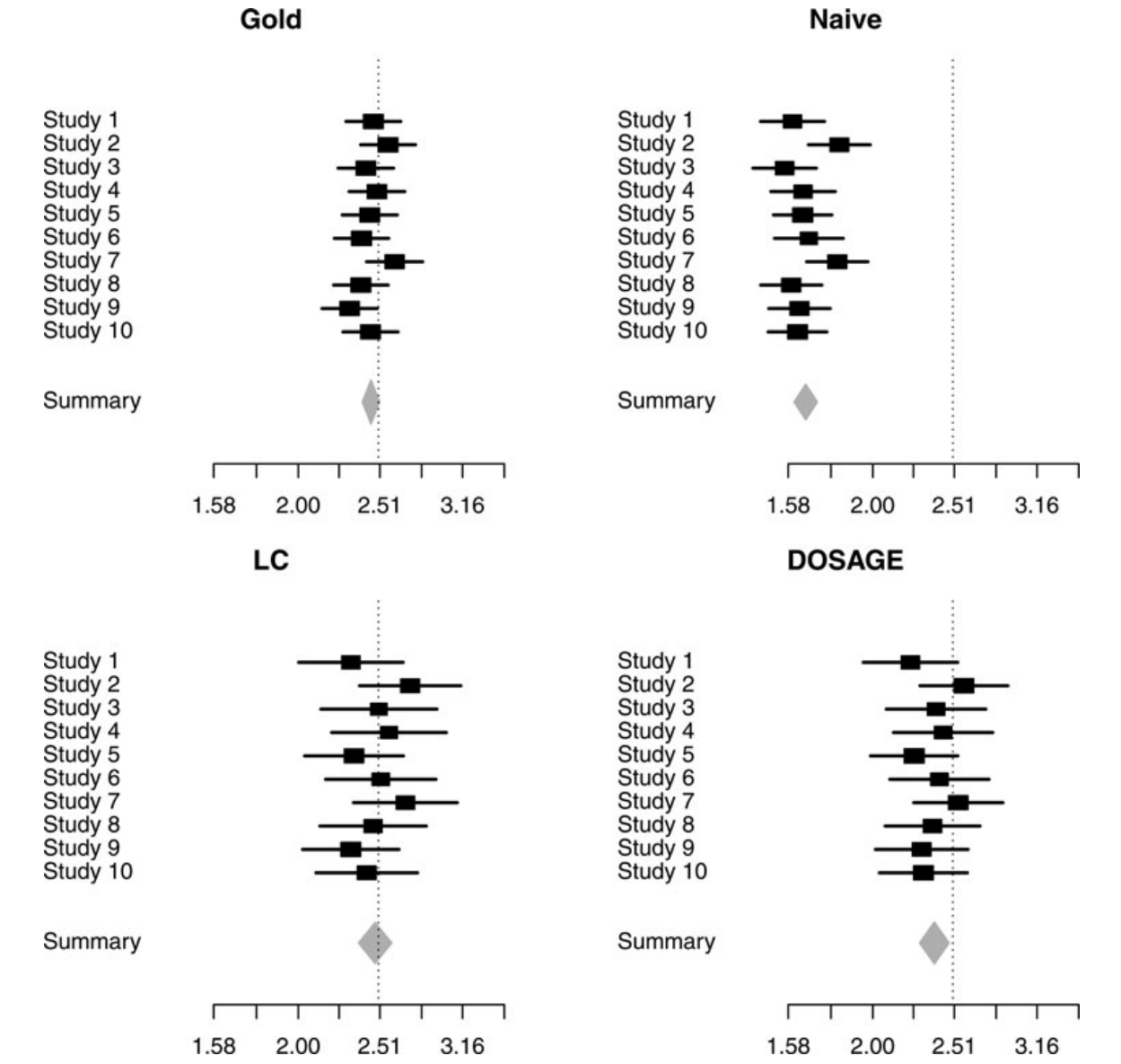


Figure 3. Meta-analysis of 10 studies, simulating censored Weibull response, a hazard ratio of 2.5, uncertainty of $R^2 = 0.3$, and a minor allele frequency = 0.5.

LC or DOSAGE approaches gave very similar results, with a pairwise Spearman correlation >99%. The correlation between the Naive approach and any other two is smaller but also high (approximately 93%).

None of the three approaches produced any significant results, except that one SNP achieved a significance level slightly beyond the Bonferroni-corrected threshold using the DOSAGE approach (see Manhattan plots represented in Figure 5). This is probably an artifact because the rest of SNPs are far from the significance level. On another hand, there seems to be no inflation of type I error rate according to QQ-plots represented in Figure 6 for none of the three approaches.

The failure rate (i.e., the number of imputed SNPs for which models have not converged have achieved a nonreliable estimation) has been very low for all three methods: <0.4%.

Computational Time

Although Aulchenko et al. [2010] cites that the LC approach is computationally demanding, observed times required to analyze Framingham-imputed SNPs on real data phenotypes did not confirm this. The number of iterations required to achieve convergence is low for the vast majority of SNPs (5

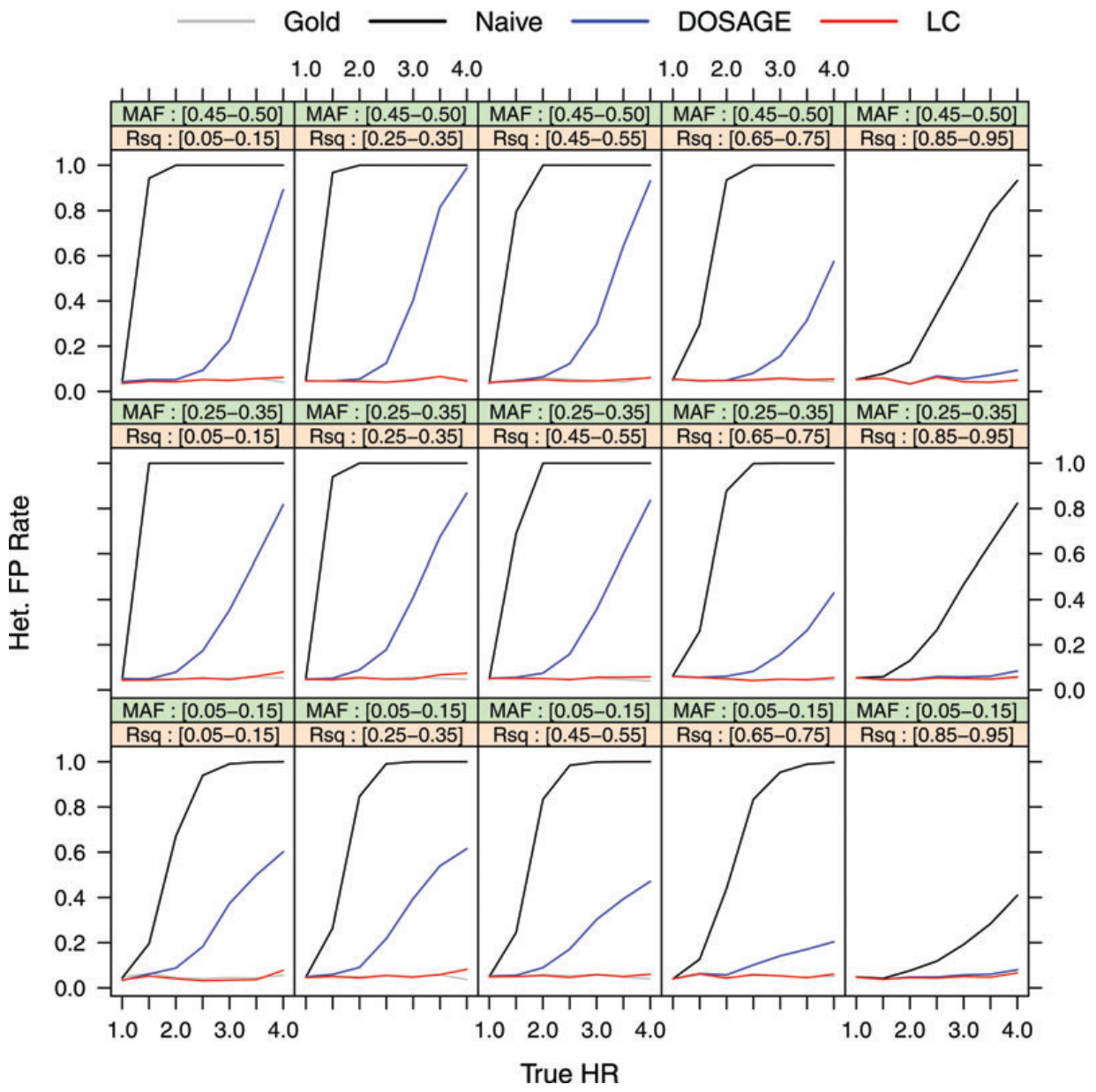


Figure 4. False positive rate (*P*-value < 0.05) of a meta-analysis taking 5 + 5 studies simulated from a censored Weibull response.

or 6 iterations). This makes the computation time consumed in fitting the LC approach model similar to that of a standard Cox regression model. To analyze 1,464,340 imputed SNPs on 3,008 individuals and adjusting by age, sex, and five genetic principal components, it took 16.7, 17.4, and 29.8 hours for Naive, DOSAGE, and LC methods, respectively, using a Linux platform x86_64-redhat-linux-gnu (64-bit) CPU with 2GHz and 16G of RAM memory, and using version 2.15.1 of R software. Therefore, we have shown that, once first and second derivatives expressions from the likelihood functions are specified, LC is not more computationally demanding than Naive or DOSAGE approaches, and it is feasible to perform GWAS analysis using the LC approach using a multicore processor; for example, with just 10 cores, analysis of a GWAS may take less than three hours.

Discussion

We studied how to incorporate uncertainty when analyzing imputed SNP in cohort design studies and conducted a comprehensive analysis comparing two existing methods with our proposed one. The simulations demonstrated that

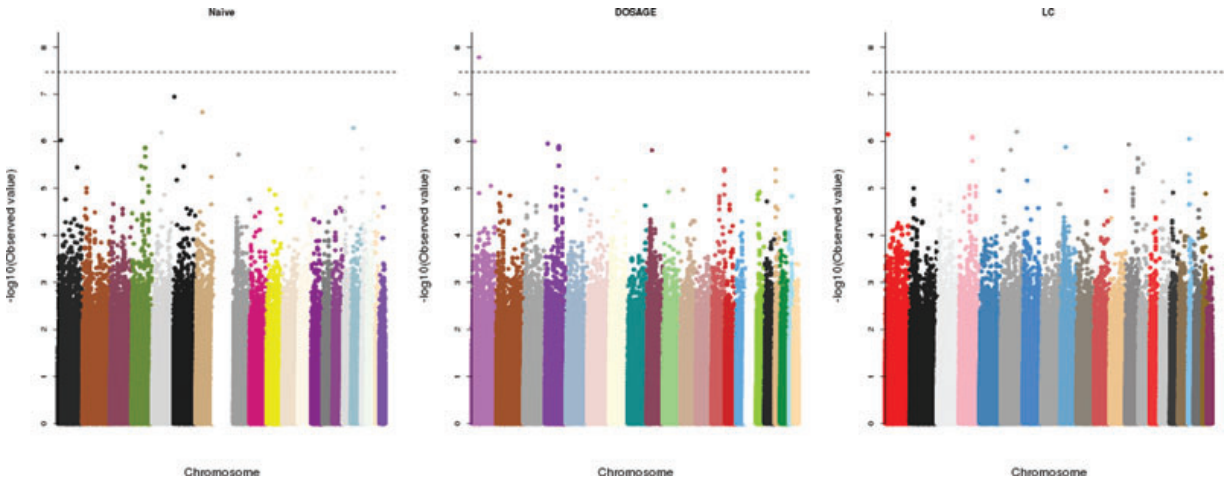


Figure 5. Manhattan plot representing minus log *P*-values of each analyzed imputed SNP from the Framingham real data.

DOSAGE is biased. This bias is negligible for small HRs, but for moderate-large effects (HR > 2 –3) the results are not correct. This may explain why DOSAGE (and other similar methods like SNPTEST) can be a good option for analyzing case-control studies, where ORs are between 1.1 and 1.5. However, longitudinal studies have shown that the effect of a given SNP may be very strong. For instance, González et al. [2011] and Frey et al. [2005] reported a HR of 3.1 and 12.1, respectively. Our proposed method, which truly incorporates uncertainty in the likelihood function properly, showed excellent results in all scenarios, in particular for moderate-high risk estimates. What is clear from our simulation studies is that the Naive method (e.g., best guess) should not be used any more, at least not when uncertainty exists.

Using real data, we also showed that none of the three methods inflates the type I error, i.e., they produce no excess of false positives. Indeed, we have shown that our proposed method takes comparable amount of time to fit the data than standard regression models such as Cox, making feasible to analyze a GWAS with 1.5 million of imputed SNPs.

Although LC and DOSAGE have very similar power for all scenarios, the latter is less accurate (it is more biased and the coverage rate is lower when the effect increases). LC and DOSAGE achieve similar power in all scenarios, although Naive is much less powerful when uncertainty increases. Good power is important to discover associated SNP. However, accurate estimations can be crucial in practical terms, such as in the case of developing a genetic risk model function.

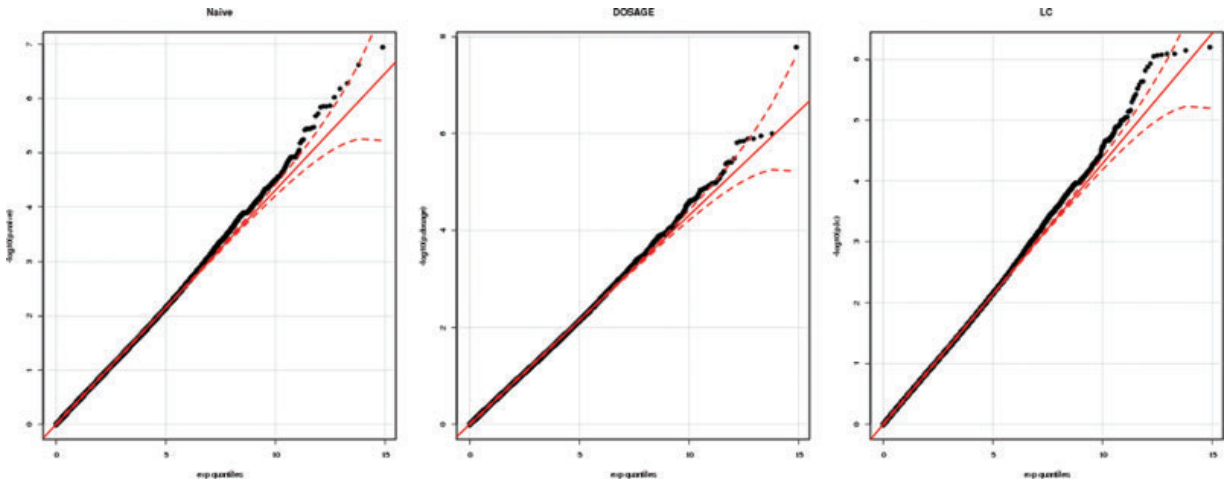


Figure 6. *P*-values Q-Q plot of each analyzed imputed SNP from the Framingham real data.

A risk model requires knowledge of the effect of a given SNP. It can be estimated using data from a single study or a better risk estimation can be obtained from a meta-analysis using published papers. In both cases a nonbiased estimation is important to achieving an accurate genetic risk model.

Our approach would also be useful in comparing the effect observed in our cohort with a published study. Let us assume that the interested SNP is not genotyped in our cohort because we use a different or less dense platform. After imputing, the LC method guarantees that both effects will be comparable; using the DOSAGE or Naive approaches may obtain a different risk estimate and we will not be able to say whether it is a real difference or is biased.

In conclusion, when analyzing longitudinal data from cohort studies, the LC approach is the only method (of the three considered) that maximizes the proper likelihood function. Therefore, it is the only one that guarantees the good properties of the maximum likelihood estimates and yields unbiased and powerful results. In practical terms, we have observed that when the effect of the SNP is high (HR > 2) only LC has fully satisfactory estimates in terms of bias, coverage, and power regardless of the degree of uncertainty. In other situations (low HRs) DOSAGE and LC behave similarly.

Acknowledgements

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Authors want to thank Elaine Lilly for her contribution in the English revision of the manuscript, and Mikel Esnaola for his great contribution in translating the algorithms code syntax to C language.

References

Alakus H, Warnecke-Eberz U, Bollschweiler E, Mönig SP, Vallböhmer D, Brabender J, Drebber U, Baldus SE, Riemann K, Siffert W, and others. 2009. GNAS1 T393C polymorphism is associated with histopathological response to neoadjuvant radiochemotherapy in esophageal cancer. *Pharmacogenom J* 9(3):202–207.

Aulchenko Y, Struchalin M, van Duijn C. 2010. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformat* 11:134.

Biernacka J, Tang R, Li J, McDonnell S, Rabe K, Sinnwell J, Rider D, Andrade M, Goode E, Fridley B. 2009. Assessment of genotype imputation methods. *BMC Proc* 3(Suppl 7):S5.

Del Greco MF, Pattaro C, Luchner A, Pichler I, Winkler T, Hicks AA, Fuchsberger C, Franke A, Melville SA, Peters A, and others. 2011. Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster. *Human Mol Genet* 20(8):1660–1671.

Fornage M, Debette S, Bis JC, Schmidt H, Ikram MA, Dufouil C, Sigurdsson T, Lumley S, De Stefano AL, Fazekas F, and others. 2011. Genome-wide association studies of cerebral white matter lesion burden. *Ann Neurol* 69(6):928–939.

Frey UH, Alakus H, Wohlschlaeger J, Schmitz KJ, Winde G, van Calker HG, Jöckel KH, Siffert W, Schmid KW. 2005. Gnas1 T393C polymorphism and survival in patients with sporadic colorectal cancer. *Clin Cancer Res* 11(14):5071–5077.

Frey UH, Lümmen G, Jäger T, Jöckel KH, Schmid KW, Rübben H, Müller N, Siffert W, Eisenhardt A. 2006a. The GNAS1 T393C polymorphism predicts survival in patients with clear cell renal cell carcinoma. *Clin Cancer Res* 12(3):759–763.

Frey UH, Nückel H, Sellmann L, Siemer D, Küppers R, Dürig J, Dührsen U, Siffert W. 2006b. The GNAS1 T393C polymorphism is associated with disease progression and survival in chronic lymphocytic leukemia. *Clin Cancer Res* 12(19):5686–5692.

Gonzalez JR, Subirana I, Escaramis G, Peraza S, Caceres A, Estivill X, Armengol L. 2009. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformat* 10: 172.

González SJ, Rojas JI, Redal MA, Patrucco L, Correale J, Argibay PF, Cristiano E. 2011. CD24 as a genetic modifier of disease progression in multiple sclerosis in argentinean patients. *J Neurol Sci* 307(1–2):18–21.

Granada M, Wilk JB, Tuzova M, Strachan DP, Weidinger S, Albrecht E, Gieger C, Heinrich J, Himes BE, Hunninghake GM, and others. 2011. A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J Allergy Clin Immunol* 129(3):840–845.

Heist RS, Zhou W, Chirieac LR, Cogan-Drew T, Liu G, Su L, Neuberg D, Lynch TJ, Wain JC, Christiani DC. 2007. MDM2 polymorphism, survival, and histology in early-stage non–small-cell lung cancer. *J Clin Oncol* 25(16):2243–2247.

Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N, and others. 2011. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 43(10):1984–1989.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 39: 906–913.

Meyre D, Delplanque J, Chèvre JC, Lecoeur C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proença C, and others. 2009. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in european populations. *Nat Genet* 41(2):157–159.

Morales E, Bustamante M, Gonzalez JR, Guxens M, Torrent M, Mendez M, Garcia-Esteban R, Julvez J, Forns J, Vrijheid M, and others. 2011. Genetic variants of the FADS gene cluster and ELOVL gene family, colostrums LC-PUFA levels, breastfeeding, and child cognition. *PloS ONE* 6(2):e17181.

Morgan T, House J, Cresci S, Jones P, Allayee H, Hazen S, Patel Y, Patel R, Eapen D, Waddy S, and others. 2011. Investigation of 95 variants identified in a genome-wide study for association with mortality after acute coronary syndrome. *BMC Med Genet* 12(1):127.

Myocardial Infarction Genetics Consortium. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet.* 41(3):334–341.

Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ, and others. 2009. Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways. *Nat Genet* 41(2):199–204.

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 42(7):570–575.

R Development Core Team. 2012. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Available at: http://www.R-project.org/. (R version 2.15.1). ISBN 3-900051-07-0.

Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, and others. 2006. App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 38(1):24–26.

Subirana I, Diaz-Uriarte R, Lucas JR, González G. 2011. CNVassoc: association analysis of CNV data using R. *BMC Med Genom* 4:47.

Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, and others. 2009. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet* 5(3):e1000433.

Trompet S, de Craen A, Postmus I, Ford I, Sattar N, Caslake M, Stott D, Buckley B, Sacks F, Devlin J, and others. 2011. Replication of LDL GWAs hits in PROSPER/PHASE as validation for future (pharmaco) genetic analyses. *BMC medical genetics* 12(1): 131.

Ueno H, Sato Y, Okusaka T, Furuse J, Ishii H, Sekine A, Nakamura Y, Kaniwa N, Sawada J, Saijo N. 2008. Association of SNPs in ABCC1 gene with overall survival in stage IV pancreatic adenocarcinoma patients treated with gemcitabine monotherapy. *J Clin Oncol* 26(15S): Abstract 14504.

Willer CJ, Li Y, Ding J, Scheet P, Abecasis GR. 2010. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34(8):816–834.

Wojnar M, Brower KJ, Strobbe S, Ilgen M, Matsumoto H, Nowosad I, Sliwerska E, Burmeister M. 2009. Association between val66met brain-derived neurotrophic factor (BDNF) gene polymorphism and post-treatment relapse in alcohol dependence. *Alcohol: Clin Exp Res* 33(4):693–702.

Xing J, Myers RE, He X, Qu F, Zhou F, Ma X, Hyslop T, Bao G, Wan S, Yang H, and others. 2011. GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis. *Eur J Cancer* 47(11):1699–1707.

Zheng J, Li Y, Abecasis GR, Scheet P. 2011. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* 35(2):102–110.

Appendix: Likelihood, Score, and Hessian Functions

The logarithm of the likelihood function (log-likelihood) for the Latent Class model is

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=0}^{2} \mathrm{P}(Y_i, \delta_i \mid \mathrm{SNP}_i = k, \boldsymbol{Z}_i; \boldsymbol{\Theta}) w_{ik} \right\},$$

where $\boldsymbol{\Theta} = (\alpha, \beta, \gamma_1, \ldots, \gamma_p, \phi)$ is the vector of parameters. For convenience, we will denote θ_s as the sth component of $\boldsymbol{\Theta}$ vector, which can be either the constant α, the SNP coefficient β, the covariates coefficients γ_j $(j = 1, \ldots, p)$, or the scalar parameter ϕ.

The first derivatives of the log-likelihood function with respect to the sth component of $\boldsymbol{\Theta}$:

$$\frac{\partial \log L(\boldsymbol{\Theta})}{\partial \theta_s} = \sum_{i=1}^{n} \frac{\sum_{k=0}^{2} \frac{\partial h_{ik}}{\partial \theta_s}}{g_i},$$

and the second derivatives of the log-likelihood function:

$$\frac{\partial^2 \log L(\boldsymbol{\Theta})}{\partial \theta_s \partial \theta_{s'}} = \sum_{i=1}^{n} \frac{\sum_{k=0}^{2} \frac{\partial h_{ik}}{\partial \theta_s \partial \theta_{s'}} g_i - \sum_{k=0}^{2} \frac{\partial h_{ik}}{\partial \theta_s} \sum_{k=0}^{2} \frac{\partial h_{ik}}{\partial \theta_{s'}}}{g_i^2},$$

where

$$h_{ik} \equiv \mathrm{P}(Y_i, \delta_i \mid \mathrm{SNP}_i = k, \boldsymbol{Z}_i; \boldsymbol{\Theta}) w_{ik},$$

$$g_i \equiv \sum_{k=0}^{2} h_{ik}.$$

Finally, we define the linear predictor as

$$\eta_{ik} \equiv \alpha + \beta k + \boldsymbol{Z}_j \boldsymbol{\gamma}',$$

where k is the number of risk alleles (i.e., 0, 1, or 2 copies), and $\boldsymbol{Z}_j$ is the covariate vector for the ith individual.

Two situations must be considered: whether the time to response for the ith individual ($\delta_i = 1$) has been observed or not ($\delta_i = 0$), where δ is called the censor indicator or variable.

If $\delta_i = 0$:

h_{ik} function takes the form:

$$h_{ik} = w_{ik} e^{-\lambda_{ik} Y_i^{\phi}},$$

where $\lambda_{ik} = \exp(\eta_{ik})$.

- First derivatives:
  – with respect to α, β, or γ_j $(\theta_s \in \{\alpha, \beta, \gamma_j\})$:

$$\frac{\partial h_{ik}}{\partial \theta_s} = h_{ik} x_{isk} (-Y_i^{\phi} \lambda_{ik}),$$

where x_{isk} takes the value of 1, k or Z_{ij} depending on whether the derivative is with respect to α, β, or γ_j, respectively, and Z_{ij} is the value of jth covariate for the ith individual.

  – with respect to the shape parameter ϕ:

$$\frac{\partial h_{ik}}{\partial \phi} = -h_{ik} \lambda_{ik} Y_i^{\phi} \log(Y_i).$$

- Second derivatives:
  – with respect to α, β, or γ_j $(\theta_s, \theta_{s'} \in \{\alpha, \beta, \gamma_j\})$:

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \theta_{s'}} = x_{isk} \left[-Y_i^{\phi} \lambda_{ik} \left(x_{is'k} h_{ik} - \frac{\partial h_{ik}}{\partial \theta_{s'}} \right) \right],$$

where x_{isk} takes the value of 1, k, or Z_{ij} depending on whether θ_s is equal to α, β, or γ_j, respectively. Similarly, $x_{is'k}$ takes the value of 1, k, or Z_{ij} depending on whether $\theta_{s'}$ is equal to α, β, or γ_j, respectively.

  – with respect to shape parameter ϕ:

$$\frac{\partial^2 h_{ik}}{\partial \phi^2} = -\lambda_{ik} \log(Y_i) Y_i^{\phi} \left(\frac{\partial h_{ik}}{\partial \phi} + h_{ik} \log(Y_i) \right).$$

- Cross-derivatives between the constant/coefficients $(\theta_s \in \{\alpha, \beta, \gamma_j\})$ and the shape parameter ϕ:

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \phi} = x_{isk} \left[-\lambda_{ik} Y_i^{\phi} \left(\frac{\partial h_{ik}}{\partial \phi} + h_{ik} \log(Y_i) \right) \right],$$

where x_{isk} takes the value of 1, k, or Z_{ij} depending on whether θ_s is equal to α, β, or γ_j, respectively.

If $\delta_i = 1$:

h_{ik} function takes the form:

$$h_{ik} = w_{ik} \lambda_{ik} \phi Y_i^{\phi - 1} e^{-\lambda_{ik} Y_i^{\phi}}.$$

- First derivatives
  – with respect to α, β, or γ_j $(\theta_s \in \{\alpha, \beta, \gamma_j\})$:

$$\frac{\partial h_{ik}}{\partial \theta_s} = h_{ik} x_{isk} (1 - Y_i^{\phi} \lambda_{ik}),$$

where x_{ij} takes the value of 1, k, or Z_{ij} depending on whether the derivative is with respect to α, β, or γ_j, respectively, and Z_{ij} is the value of jth covariate for the ith individual.

  – with respect to the shape parameter ϕ:

$$\frac{\partial h_{ik}}{\partial \phi} = -h_{ik} \left(\lambda_{ik} Y_i^{\phi} \log(Y_i) - \log(Y_i) - \frac{1}{\phi} \right).$$

- Second derivatives:
  – with respect to α, β, or γ_j $(\theta_s, \theta_{s'} \in \{\alpha, \beta, \gamma_j\})$:

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \theta_{s'}} = x_{isk} \left[\frac{\partial h_{ik}}{\partial \theta_{s'}} \left(1 - \lambda_{ik} Y_i^{\phi} \right) - h_{ik} \lambda_{ik} Y_i^{\phi} x_{is'k} \right],$$

where x_{isk} takes the value of 1, k, or Z_{ij} depending on whether θ_s is equal to α, β, or γ_j, respectively. Similarly, $x_{is'k}$ takes the value of 1, k, or Z_{ij} depending on whether $\theta_{s'}$ is equal to α, β, or γ_j, respectively.

– with respect to shape parameter ϕ:

$$\frac{\partial^2 h_{ik}}{\partial \phi^2} = -\frac{\partial h_{ik}}{\partial \phi}\left(Y_i^{\phi}\log(Y_i)\lambda_{ik} - \log(Y_i) - \frac{1}{\phi}\right) - h_{ik}\left(Y_i^{\phi}\log^2(Y_i)\lambda_{ik} + \frac{1}{\phi^2}\right).$$

- Cross-derivatives between the constant/coefficients ($\theta_s \in \{\alpha, \beta, \gamma_j\}$) and the shape parameter ϕ:

$$\frac{\partial^2 h_{ik}}{\partial \theta_s \partial \phi} = x_{isk}\left[\frac{\partial h_{ik}}{\partial \phi}\left(1 - \lambda_{ik}Y_i^{\phi}\right) - h_{ik}\lambda_{ik}Y_i^{\phi}\log(Y_i)\right],$$

where x_{isk} takes the value of 1, k, or Z_{ij} depending on whether θ_s is equal to α, β, or γ_j respectively.

Subirana et al. BMC Medical Genomics 2011, 4:47
http://www.biomedcentral.com/1755-8794/4/47

BMC Medical Genomics

SOFTWARE Open Access

CNVassoc: Association analysis of CNV data using R

Isaac Subirana[1,2,3], Ramon Diaz-Uriarte[4], Gavin Lucas[2] and Juan R Gonzalez[5,1*]

Abstract

Background: Copy number variants (CNV) are a potentially important component of the genetic contribution to risk of common complex diseases. Analysis of the association between CNVs and disease requires that uncertainty in CNV copy-number calls, which can be substantial, be taken into account; failure to consider this uncertainty can lead to biased results. Therefore, there is a need to develop and use appropriate statistical tools. To address this issue, we have developed `CNVassoc`, an R package for carrying out association analysis of common copy number variants in population-based studies. This package includes functions for testing for association with different classes of response variables (e.g. class status, censored data, counts) under a series of study designs (case-control, cohort, etc) and inheritance models, adjusting for covariates. The package includes functions for inferring copy number (CNV genotype calling), but can also accept copy number data generated by other algorithms (e.g. CANARY, CGHcall, IMPUTE).

Results: Here we present a new R package, CNVassoc, that can deal with different types of CNV arising from different platforms such as MLPA o aCGH. Through a real data example we illustrate that our method is able to incorporate uncertainty in the association process. We also show how our package can also be useful when analyzing imputed data when analyzing imputed SNPs. Through a simulation study we show that CNVassoc outperforms CNVtools in terms of computing time as well as in convergence failure rate.

Conclusions: We provide a package that outperforms the existing ones in terms of modelling flexibility, power, convergence rate, ease of covariate adjustment, and requirements for sample size and signal quality. Therefore, we offer CNVassoc as a method for routine use in CNV association studies.

Background

The proportion of variation in risk of complex diseases explained by the single nucleotide polymorphisms (SNPs) that have been discovered in recent years using the genome-wide association approach appears to limited. This has lead to the suggestion that other, possibly more complex, genetic variants could partly explain the remaining disease susceptibility. Technological advances now allow a class of genetic variants known as copy number variants (CNV) to be genotyped with increasing levels of accuracy, and several studies have recently explored the relationship between these variants and risk of complex disease [1,2]. Genotyping these kinds of complex genetic markers is still a challenge and current laboratory techniques and platforms often contain a non-negligible percentage of errors. In order to minimise bias in the results of association studies involving CNVs, uncertainty in these copy number calls must be taken into account in the analysis. In addition, large-scale CNV genotyping projects need a tool to automate the analysis of thousands of CNVs. Here, we present `CNVassoc`, an R package [3] designed to analyze CNV data. Methodological details of the algorithms and applications implemented in `CNVassoc` are described in [4]. In addition to these, other techniques, such as accounting for batch effects in inferring copy number status, or modelling other response distributions (Poisson or Weibull for censored data) have now been incorporated into `CNVassoc`. In this application note we present an overview of the package. The Additional file 1 contains a tutorial (the vignette for the package) together with technical notes on the derivation of the likelihoods for the different models.

Implementation

We developed a set of functions to analyse copy number variants and integrated them as an R package called

* Correspondence: jrgonzalez@creal.cat
[5]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
Full list of author information is available at the end of the article

Subirana et al. BMC Medical Genomics 2011, 4:47
http://www.biomedcentral.com/1755-8794/4/47

CNVassoc. Also, we created a very extensive manual of the package (vignette) with several examples of real and simulated data explaining how to use the package functions and their capabilities.

The R software is a general purpose and open source program commonly used in all type of statistical analysis. Having incorporated the functions as an R package allows user to take advantage of R flexibility in manipulating the input and the results when analysing CNVs with CNVassoc. In addition, we structured CNVassoc functions and results in methods and classes to make the package usage easier and more intuitive.

Software main features

To date, only one other R package, CNVtools[5], has been developed that can appropriately incorporate CNV copy number call uncertainty in the test for association between CNVs and disease. However, CNVtools has some limitations, mainly related to the fact that the copy number calling and association testing steps are combined in a single procedure. The current version of CNVtoolshttp://bioconductor.org uses complex and computationally intensive algorithms, cannot adjust for covariates, and can only model binary and normally distributed responses. By separating these two steps, CNVassoc offers significant advances in terms of analytical flexibility and computational speed.

Inferring copy number status

By separating the CNV calling and association testing steps, CNVassoc allows the user to test for association between CNVs and disease using copy number probabilities from any source. While the use of probability data from more powerful calling algorithms such as CGHcall [6], IMPUTE [7,8] or CANARY [9] is recommended, CNVassoc provides several tools for inferring copy number status, where necessary. For example, CNVassoc can fit a mixture of normal distributions to CNV signal intensity data [10], or assign copy number status by defining a set of signal intensity cut points, which might be useful when analysing probe intensity data from MLPA [11] or qPCR [12]. In addition, there is an option to take batch effects into account, in order to reduce false positives and provide robust estimates, as discussed in [5].

Considering batch effect

In CNVassoc, the batch effect has been handled in the following way:

Formally, the intensity signal distribution, y, is supposed to follow a mixture of gaussian distributions,

$$f(y_{ib}) = \sum_c \phi(\mu_{cb}, \sigma_{kb}) w_c$$

where, φ is the gaussian density function, μ_{cb} and σ_{cb} is the mean and standard deviation respectively of intensity signal for c copy number variants in b-th batch, and w_c is the proportion of individuals with c copies in the population. Notice that mean and standard deviation can vary not only between copy number status but also between batches, but the copy number status prevalences (w_c) not. If μ_{cb} and σ_{cb} varies between batches and batches are associated with the disease/response, then the batch effect exists by definition, and can lead to false association if it is not taken into account [5].

In CNVassoc, specific means, standard deviations and prevalences estimates are calculated separately using data from each batch. Then, prevalences estimates are obtained averaging from specific prevalences:

$$\hat{w}_c = \sum_{b=1}^{B} n_b \hat{w}_{cb}/n$$

where n_b is the number of sample individuals in the b-th batch, B is the total number of batches in the sample, and n is the total number of individuals in the sample.

Improved association test

To incorporate CNV copy number uncertainty in the association test, CNVassoc uses a simpler model formulation than that of CNVtools. This allows us to use the faster Newton-Raphson procedure, which yields not only the effect estimate for the CNV, but also its confidence interval.

Adjustment for covariates

CNVassoc can fit association models adjusted for covariates (age, gender, smoking, etc.), which may be particularly important where it is necessary to adjust for population stratification [13].

Response phenotypes

CNVassoc can be used to analyse dichotomous (Binomial), count (Poisson), or continuous (Gaussian) response phenotypes, as well as data from cohort studies (Weibull).

Inheritance models

CNVassoc can perform association analysis under a codominant (additive) model, which assumes a constant effect on phenotype per unit change in copy number, or under a model-free design, which treats each copy number as an independent category.

Analysis of multiple CNVs

To perform association testing of multiple CNVs with greater computational efficiency, a function called

Subirana *et al. BMC Medical Genomics* 2011, **4**:47
http://www.biomedcentral.com/1755-8794/4/47

multiCNVassoc has been implemented. When multiple processors are available, it can parallelize association tests using the Snow package http://www.sfu.ca/~sblay/R/snow.html. An example of association tests involving several CNVs is shown in Section 3 of the Additional file 1 where data from a CGH array is analysed.

Computational Efficiency

Using the same sample sizes and probe signal intensity distributions as used in [5], we performed a simulation study in order to compare the performance of the methods implemented in CNVassoc and CNVtools. We observed that both methods performed well, but we note that CNVassoc has a number of important advantages over CNVtools in terms of computational speed and robustness in situations of limited sample sizes.

Performing association tests

First, an object of class cnv must be created by CNVassoc or using probabilities from other algorithms. Then, an association test between the CNV and disease can be performed using the CNVassoc function, which returns an object of class 'CNVassoc'. Associated print and summary functions give exhaustive outputs. The (CNVtest) function computes an overall p-value to test whether a CNV is associated with the disease

Functions to simulate CNV data

In CNVassoc package, function to simulate CNV data have been implemented. It is possible to simulate data from different type of responses and studies: case-control (simCNVdataCaseCon), cohort with binary response (simCNVdataBinary), counting process with poisson-distributed response (simCNVdataPois), quantitative normal-distributed response (simCNVdataNorm) and time-to-event with right-censored-weibull-distributed response (simCNVdataWeibull).

Association analysis on imputed SNPs

Also, it is possible to analyse association of imputed SNPs and response. Taking the genotypes probabilities obtained from any software capable to impute SNPs, such as IMPUTE [7,8], association analysis for case-control studies, cohort, quantitative or counting response can be performed with CNVassoc. In section 5 of the Additional file 1 we show in detail how to analyse a data set downloadable from SNPTEST website which contains probabilities of different imputed genotypes from different SNPs among a set of cases and controls.

Results and Discussion

In this section we show the results obtained in inferring copy number status and association analysis on a real data set including 360 cases and 291 controls (data described in [4]). The data contains peaks intensities for two genes arising from an MLPA assay. From this example, we present the main CNVassoc functions and illustrate how to use them to infer copy number copies and estimate association on case-control status.

A more detailed description of all these analyses and others (imputed SNPs, aCGH data, other phenotypes distributions -poisson, weibull and normal-) can be found in Additional file 1.

Inferring copy number status

Previous to association analysis, inferring copy number status process must be done. To do so, the function cnv is used. In this subsection, gene 2 from MLPA data example is used. This data set can be load from the CNVassoc package.

>library(CNVassoc)

>data(dataMLPA)

>CNV <- cnv(x = dataMLPA$Gene2, threshold.0 = 0.01, mix.method = "mixdist")

The peak intensities of gene 2 are assumed to follow a mixture of normal distributions, and the method used to estimate this distribution is specified by the mix.method argument. When threshold.0 = 0.01, all individuals with peak intensities lower than 0.01 are assumed to carry 0 copies. The CNV object is of class cnv, which can be printed and plotted (Figure 1).

```
>CNV
Inferred copy number variant by a quantitative signal
  Method: function mix {package: mixdist}
```

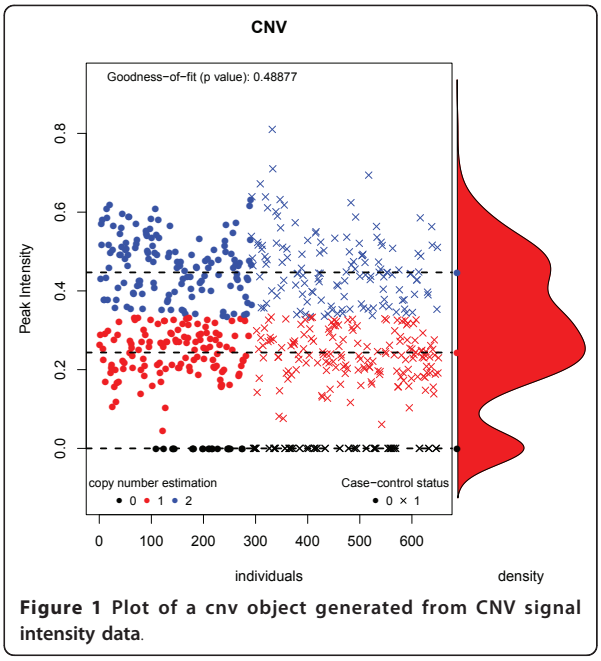


Figure 1 Plot of a cnv object generated from CNV signal intensity data.

```
 -. Number of individuals: 651
 -. Copies 0, 1, 2
 -. Estimated means: 0, 0.2435, 0.4469
 -. Estimated variances: 0, 0.0041, 0.0095
 -. Estimated proportions: 0.1306, 0.4187, 0.4507
 -. Goodness-of-fit test: p-value = 0.4887659
 -. Note: number of classes has been selected using the best BIC
```

>plot(CNV)

A measure that quantifies the amount of uncertainty in the CNV calling estimation can be computed using the function getQualityScore. Various measures are available; the following is an example of how to obtain the quality score (uncertainty measure) described in the CNVtools paper [5]:

>getQualityScore(CNV, type = "CNVtools")

```
–CNVtools Quality Score: 3.057171
```

In some cases, it may be preferable to infer copy number status using another algorithm that is not implemented in CNVassoc, e.g. if the probe signal intensities do not follow a mixture of normal distributions. A matrix of copy number probabilities obtained from other algorithms can be used as input for the cnv function to create a cnv class object, which can then be used to perform association analysis. Also, it is possible to take suspected batch effects in the signal intensity distributions into account by specifying the batch variable using the batch argument in the cnv function. This is important in order to avoid false positives in the posterior association model estimation, as suggested in [5]. A more detailed explanation and example of this issue can be found in section 4.2 of Additional File 1.

Performing association models

To carry out association analysis between CNV and disease, the function CNVassoc is used. This function incorporates copy number call uncertainty by using a latent class model as described in [4]. The response variable (disease) can be: binary, quantitative (normally distributed), from a counting process, time to event (Weibull distributed). Also, an additive or model-free pattern of inheritance can be analysed. The result returned by the CNVassoc function is an object that can be printed and summarized and its structure is very similar to other well known R functions such as glm.

Here, we continue with the same MLPA data taking the CNV object for gene 2 in the previous section. To fit a logistic regression model with case-control status as a response and CNV copy number as a predictor, and assuming an additive genetic effect, we type

>mod <- CNVassoc(casco ~ CNV, data = dataMLPA)

>summary(mod)

```
 Call:
 CNVassoc(formula = casco ~ CNV, data = dataMLPA)
 Deviance: 876.396
 Number of parameters: 3
 Number of individuals: 651
 Coefficients:
    OR lower.lim upper.lim SE Stat pvalue
 CNV0  1.0000
 CNV1  0.4772  0.2742  0.8304  0.2827 -2.6172   0.009
 CNV2  0.3169  0.1834  0.5477  0.2791 -4.1169  3.84e-05
 (Dispersion parameter for binomial family taken to be 1)
 Covariance between coefficients:
    CNV0   CNV1   CNV2
 CNV0  0.0613 0.0000  0.0000
 CNV1       0.0186 -0.0032
 CNV2            0.0166
```

By applying the summary function to the result, we obtain odds ratios, confidence intervals, and p-values for every copy number status with respect to the reference copy number category.

To compute the global CNV significance p-value, the CNVtest function can be used as follows:

>CNVtest(mod, "LRT")

```
—CNV Likelihood Ratio Test—
 Chi = 18.75453 (df = 2), pvalue = 8.462633e-05
```

In this example, a Likelihood Ratio Test (LRT) is computed, comparing a model containing CNV to a model lacking CNV (i.e. a model without predictors or the null model).

Using the CNVassoc function it is possible to change the inheritance model to additive (changing the model argument), or adjust for other covariates (such as age, sex, or principal components) in the formula argument in the usual way. Also, other types of response can be analysed changing the family argument. More detailed examples are in the Additional file 1.

Response phenotypes: Weibull

In this section, we illustrate how to analyse a time-to-event response variable (Weibull distributed) using simulated data generated with the function simCNVdataWeibull. In the following example, a CNV has been generated with 0, 1 and 2 possible copies with probabilities of 25%, 50% and 25% respectively, with intensity signal standard deviation of 0.4 for each copy status, and means of 0, 1 and 2 respectively. The response variable has been simulated under a Weibull distribution with shape parameter equal to 1 and disease incidence equal to 0.05 (per person-year) among the population

Subirana et al. BMC Medical Genomics 2011, 4:47
http://www.biomedcentral.com/1755-8794/4/47

with zero copies (reference). The proportion of observed events (non-censored) was set to 10%. Finally, these data have been generated assuming a additive CNV effect with a Hazard Ratio of 1.5 per copy.

```
>set.seed(123456)
>n <- 5000
>w <- c(0.25, 0.5, 0.25)
>mu.surrog <- 0:2
>sd.surrog <- rep(0.4, 3)
>hr <- 1.5
>incid0 <- 0.05
>lambda <- c(incid0, incid0 * hr, incid0 * hr^2)
>shape <- 1
>scale <- lambda^(-1/shape)
>perc.obs <- 0.1
>time.cens <- qweibull(perc.obs, mean(shape), mean(scale))
>dsim <- simCNVdataWeibull(n, mu.surrog, sd.surrog, w, lambda,
+   shape, time.cens)
```

Once the CNV data and phenotype has been generated, inferring copy number status and fitting the association model is performed in the following two steps:

(1) Inferring copy number status, as for case-control studies:

```
>CNV <- cnv(dsim$surrog, mix = "mclust")
>attr(CNV, "num.copies") <- 0:2
```

Note that 3 copy number statuses has been estimated by BIC criteria. By default 1, 2 and 3 copies are assigned. The number of copies for each status can be changed to 0, 1 and 2 respectively by modifying the num.copies attribute.

2) Testing for association between CNV and time-to-event, specifying the family argument as "weibull":

```
>fit <- CNVassoc(Surv(resp, cens) ~ CNV, data = dsim, family = "weibull",
+   model = "add")
>coef(summary(fit))
```

```
                    HR lower.lim  upper.lim     SE   stat    pvalue
    trend  1.385556  1.205619  1.592348 -0.07097498  4.594595  4.335896e-06
```

Note that, Hazard Ratios (HR) are displayed instead of Odds Ratios. In this case, an additive CNV effect has been assumed in performing the association model.

Computational Efficiency

In this section, we compare the performance of `CNVassoc` in terms of speed and convergence rate to that of `CNVtools`, which is the only other tool that is currently available for performing CNV association analysis, while correctly taking copy number uncertainty into account. Simulated case-control data was generated for different sample sizes (500 cases and 500 controls; 2,000 cases and 2,000 controls), and different degrees of call uncertainty, from very little uncertainty (Q = 6) to a moderate-high degree of uncertainty (Q = 3). A single CNV marker has been simulated using 1,000 iterations (simulations), under the described scenarios. In each simulation, univariate probe signal intensities (similar to MLPA) have been generated from a gaussian mixture distribution, and copy number status has been inferred from them. After this, an association model has been performed using the proposed method (Latent Class model). The uncertainty measure, Q, was proposed by [5] (see page 3); values of Q below 3 indicate moderate-high uncertainty and this must be taking into account in the association analysis, while values of Q bigger than 4.5 or 5 indicate that uncertainty is almost insignificant. Table 1 shows the number of times model estimation fails using `CNVassoc` and `CNVtools` under these various scenarios. `CNVassoc` converges in all simulations, except when sample size is small and uncertainty is high. When sample size is high (2,000 cases and 2,000 controls) `CNVassoc` converges in all situations, while `CNVtools` fails in some simulations when uncertainty is high. And when sample size is moderate-low (500 cases and 500 controls), `CNVassoc` converges almost in all times except when uncertainty is high (Q < 3.5), while `CNVtools` fails in some simulations even when the degree of uncertainty is low (Q = 6) and starts to fail in the majority of situations when uncertainty is moderate (Q < 4) and performs even worse when is high.

We have also observed a marked difference in the speed of each procedure: when analyzing 10,000 CNVs in 2,000 cases and 2,000 controls, and with a Q = 4, `CNVtools` took 1 day and 17 hours to complete the

Table 1 Number of failed convergence simulations out of 500 using CNVassoc and CNVtools according to inferring copy number uncertainty *Q* and number of cases *N*

	N = 2000		*N* = 500	
Q	CNVassoc	CNVtools	CNVassoc	CNVtools
6.0	0	0	0	15
5.5	0	0	0	20
5.0	0	0	0	65
4.5	0	0	0	92
4.2	0	0	0	187
4.0	0	0	0	246
3.7	0	0	0	294
3.5	0	1	0	299
3.2	0	13	212	389
3.0	0	65	331	400

analysis, whereas CNVassoc took just 90 minutes; with $Q = 3$, CNVtools took 6 days and 16 hours, but CNVassoc took only 2 hours. More comparisons between CNVassoc and CNVtools are shown in section 4.3.1 of Additional file 1.

Conclusions

We present a new package for performing analysis of association between copy number variants and disease, appropriately taking uncertainty in CNV copy number calls into account. The numerical procedure for fitting the model is simple and computationally efficient, handling thousands of CNVs in reasonable time. In addition, it is possible to adjust for covariates which may be necessary to control for population stratification. A central feature of CNVassoc is that input data can come from any CNV calling algorithm that produces copy number probabilities. Note that the CNVassoc package can also be applied to SNPs. For instance, in the context of imputed SNPs (e.g., IMPUTE [7,8], BIMBAM [14], MACH1 http://www.sph.umich.edu/csg/abecasis/MACH/, etc.) the probability estimates of each genotype coming from this software can easily be incorporated to our functions. We intend to continue developing the package, and expect to incorporate CNV * non-genetic predictor interactions, and CNV * CNV interactions, in the near future.

In conclusion, considering the advantages in terms of modelling flexibility, power, convergence rate, ease of covariate adjustment, and requirements for sample size and signal quality, we offer CNVassoc as a method for routine use in CNV association studies.

Availability and requirements

1. Project name: CNVassoc
2. Project home page: http://www.creal.cat/jrgonzalez/software.htm and http://www.cran.r-project.org
3. Operating system(s): Platform independent
4. Programming language: R
5. R Dependencies: mixdist, mclust, survival
6. R Suggested: CGHcall, CGHregions, snow, CNVtools
7. License: GPL or newer

Additional material

Additional file 1: User's manual. CNVassoc_manual.pdf is the user's guide of CNVassoc package, where detailed examples with real and simulated data are shown, illustrating how to use the CNVassoc package functions.

Acknowledgements

The authors would like to express their gratitude to Dave MacFarlane and Alejandro Caceres for their helpful comments and for reviewing the manuscript. This work has been supported by the Spanish Ministry of Science and Innovation (MTM2008-02457 to JRG, BIO2009-12458 to RD-U and statistical genetics network MTM2010-09526-E (subprograma MTM) to JRG, IS, GL and RD-U). GL is supported by the Juan de la Cierva Program of the Spanish Ministry of Science and Innovation.

Author details

[1]CIBER Epidemiology and Public Health (CIBERESP), Barcelona, Spain. [2]Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, Institut Municipal d'Investigaci/'o Mèdica (IMIM), Barcelona, Spain. [3]Statistics Department, University of Barcelona (UB), Barcelona, Spain. [4]Structural Biology and Biocomputing Programme, Spanish National Cancer Centre (CNIO), Madrid, Spain. [5]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

Authors' contributions

JRG conceived the idea of incorporation probabilities to address uncertainty in CNV association studies. IS and JRG created the R functions and the package. IS implemented some R functions to simulate CNV data. GL drafted the manuscript. IS, GL, RD-U and JRG designed, performed and interpreted the simulation studies to compare CNVtools and CNVassoc. IS, RD-U and JRG helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 23 December 2010 Accepted: 24 May 2011
Published: 24 May 2011

References

1. E Gonzalez, H Kulkarni, H Bolivar, A Mangano, R Sanchez, G Catano, RJ Nibbs, BI Freedman, MP Quinones, MJ Bamshad, KK Murthy, BH Rovin, W Bradley, RA Clark, SA Anderson, RJ O'Connell, BK Agan, SS Ahuja, R Bologna, L Sen, MJ Dolan, SK Ahuja, The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science. **307**(5714):1434–40 (2005). doi:10.1126/science.1101160
2. C Le Marechal, E Masson, JM Chen, F Morel, P Ruszniewski, P Levy, C Ferec, Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet. **38**(12):1372–4 (2006). doi:10.1038/ng1904
3. R Development Core Team, *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2008) http://www.R-project.org. [ISBN 3-900051-07-0]
4. JR Gonzalez, I Subirana, G Escaramis, S Peraza, A Caceres, X Estivill, L Armengol, Accounting for uncertainty when assessing association between copy number and disease: a latent class model. BMC Bioinformatics. **10**, 172 (2009). doi:10.1186/1471-2105-10-172
5. C Barnes, V Plagnol, T Fitzgerald, R Redon, J Marchini, D Clayton, ME Hurles, A robust statistical method for case-control association testing with Copy Number Variation. Nature Genetics. **40**(10):1245–52 http://cnv-tools.sourceforge.net/CNVtools.html (2008). doi:10.1038/ng.206
6. MA van de Wiel, KI Kim, SJ Vosse, WN van Wieringen, SM Wilting, B Ylstra, CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics. **23**(7):892–894 (2007). doi:10.1093/bioinformatics/btm030
7. J Marchini, B Howie, S Myers, G McVean, P Donnelly, A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics. **39**, 906–913 (2007). doi:10.1038/ng2088
8. BN Howie, P Donnelly, J Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics. **6** (2009)
9. JM Korn, FG Kuruvilla, SA McCarroll, A Wysoker, J Nemesh, S Cawley, E Hubbell, J Veitch, PJ Collins, K Darvishi, C Lee, MM Nizzari, SB Gabriel, S Purcell, MJ Daly, D Altshuler, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. **40**(10):1253–60 (2008). doi:10.1038/ng.237
10. J Du, Combined Algorithms for Fitting Finite Mixture Distributions. *PhD thesis*. (McMaster University, Ontario, Canada, 2002)
11. JP Schouten, CJ McElgunn, R Waaijer, D Zwijnenburg, F Diepvens, P G, Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. **30**(12):e57 (2002). doi:10.1093/nar/gnf056
12. J Hellemans, G Mortier, A De Paepe, F Speleman, J Vandesompele, qBase relative quantification framework and software for management and

automated analysis of real-time quantitative PCR data. Genome Biol. **8**(2): R19 (2007). doi:10.1186/gb-2007-8-2-r19
13. A Caceres, X Basagaña, J Gonzalez, Multiple correspondence discriminant analysis: An application to detect stratification in copy number variation. Stat Med. **29**(10):3284–93 (2010)
14. B Servin, M Stephens, Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. **3**(7):e114 (2007)

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1755-8794/4/47/prepub

doi:10.1186/1755-8794-4-47
Cite this article as: Subirana *et al.*: CNVassoc: Association analysis of CNV data using R. *BMC Medical Genomics* 2011 **4**:47.

5.2. Autoría

Isaac Subirana cumple todos los criterios de autoría según establece el Comité Internacional de Editores en Revistas Médicas (http://www.icmje.org). Su grado de independencia ha ido aumentando a medida que avanzaba la tesis hasta llegar a ser autónomo como investigador científico, hecho que ha sido plasmado con la escritura de nuestro último artículo científico que se encuentra actualmente en revisión en una revista Internacional. En los dos últimos artículos, los únicos co-autores son el doctorando y el director de esta tesis, quedando de esta forma demostrado su grado de autoría. En los otros dos artículos ha participado en la redacción del artículo, en el desarrollo metodológico siguiendo las ideas que hemos discutido conjuntamente, así como en el análisis de datos e implementación del paquete de R que se ha creado para implementar toda la metodología estadística propuesta en esta tesis doctoral, tal y como queda reflejado en en el apartado de "Author's contribution".

Ninguno de los artículos científicos incluidos en esta tesis doctoral forman parte de otra tesis o ningún otro trabajo académico.

Firmado por el director de la tesis:

Juan Ramón González Ruiz

5.3. Resumen de las publicaciones

5.3.1. Artículo 1: *Accounting for uncertainty when assessing association between copy number and disease: a latent class model.*

Métodos

En este articulo se ha presentado el método estadístico propuesto en esta tesis, que se ha denominado *Latent Class*, para analizar los estudios de asociación entre los CNVs en los diseños de tipo caso-control y en los de respuesta cuantitativa. El método propuesto se ha comparado con dos de los métodos más usados en el estudio de asociación de CNVs: *threshold* y Naive.

A diferencia del *Latent Class*, ni el método *threshold* ni el método Naive tienen en cuenta la posible incertidumbre en los CNVs. La característica común de ambos es que imputan el número de copias a partir de las intensidades para posteriormente ajustar un modelo de regresión lineal o regresión logística para respuesta cuantitativa o binaria respectivamente, o un test χ^2 o de Fisher cuando no se incorporan covariables. Esto es, como si el número de copias asignado fuera el correcto para todos los individuos de la muestra. La diferencia recae en la manera que tienen de asignar el número de copias: mientras que el método *threshold* usa puntos de corte fijados a priori y a menudo arbitrarios, el método Naive usa puntos de corte a partir de métodos estadisticos de *clustering* y por lo tanto no tan "subjetivos". No obstante, es conocido que no tener en cuenta la posible incertidumbre en los CNVs conlleva estimaciones sesgadas y una falta de potencia estadística en estimar la asociación entre los CNVs y la respuesta.

Así pues, el objetivo de este trabajo ha sido averiguar bajo qué escenarios, según el grado de incertidumbre en los CNVs y el tipo de variable respuesta (binaria o cuantitiva), etc., los métodos *threshold* y Naive están "demasiado" sesgados, confirmar que el método propuesto (*Latent Class*) da estimaciones insesgadas, y compararlo con los otros dos en cuanto a potencia estadística (probabilidad de detectar CNVs asociados).

En primer lugar se ha descrito formalmente el modelo propuesto. En segundo lugar, se han mencionado los distintos métodos que proveen de las probabilidades de tener

k número de copias para cada individuo de la muestra y que son necesarias para posteriormente poder ajustar el modelo *Latent Class* y estimar la asociación entre el CNV y la variable respuesta. Por último, y para llevar a cabo la evaluación del modelo propuesto y su comparación con los métodos existentes, se ha realizado un estudio con datos simulados bajo distintos grados de incertidumbre y magnitud de asociación. Esto ha servido para cuantificar el sesgo, la precisión y la potencia estadística de cada método, bajo cada escenario simulado. Además, se ha analizado una muestra de datos reales consistente en un *array* CGH (metodología para medir CNVs a lo largo del genoma).

Resultados

A partir del estudio con datos simulados, se ha visto como tanto el método *threshold* como el Naive dan estimaciones negativamente sesgadas en todos los escenarios, mientras que el método propuesto es insesgado en todos ellos. Además, la potencia estadística alcanzada con el método propuesto ha sido mucho mayor que en los otros dos, siendo el método Naive más potente que el método *threshold*.

Para el estudio con datos reales, se han analizado dos CNVs, ambos medidos bajo una técnica mucho más precisa y cara (PCR) con lo que prácticamente no se tiene incertidumbre, y al mismo tiempo se ha medido la intensidad de la señal. Esta última técnica, en cambio, al no medir el CNV directamente, conduce a una mayor incertidumbre. Los resultados, comparando ambas técnicas, han indicado que las estimaciones obtenidas con el método *Latent Class* a partir de las intensidades es muy similar a la obtenida del CNV medido sin incertidumbre.

En resumen, mediante el método propuesto se han obtenido resultados muy satisfactorios en el análisis de estudios de asociación de CNVs, aunque éstos estén medidos con alto grado de incertidumbre, tanto en diseños de caso-control como de respuesta continua. Por otro lado, los métodos "clásicos" comúnmente usados no son fiables para en el análisis de CNVs con cierto grado de incertidumbre. En la práctica, esto se traduce en descartar gran parte de CNVs considerados en un estudio de asociación, mientras que con el método propuesto se pueden analizar todos ellos de forma óptima.

5.3.2. Artículo 2: *Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies.*

Métodos

En este artículo se ha estudiado el papel que juega la incertidumbre en los SNPs imputados en los estudios de asociación. Para ello, se ha aplicado el modelo propuesto presentado en el primer artículo para los SNPs imputados en lugar de los CNVs. Además se ha adaptado la ecuación desarrollada para estudios de casos y controles o para respuesta cuantitativa a estudios de cohorte con seguimiento (longitudinales).

En los últimos años, y gracias al proyecto HapMap, cada vez se imputan un mayor número de SNPs. Paralelamente, algunos estudios de cohorte han reportado un gran efecto entre algunos SNPs y enfermedades comunes tales como el cáncer, el Alzheimer o patologías cardiovasculares. Por ello, es de gran importancia evaluar las herramientas usadas para estimar la asociación de SNPs imputados y la enfermedad en estudios de cohorte. Debido a que las enfermedades comunes tienen una incidencia pequeña en la población, los GWAS con datos longitudinales requieren de un gran tamaño de muestra. Por este motivo se suelen juntar diferentes muestras para posteriormente meta-analizar sus resultados (*consortiums*).

El objetivo de este artículo ha sido desarrollar modelos estadísticos para SNPs imputados en estudios longitudinales, ampliando el método presentado en el primer artículo a este tipo de variante genética y a estudios con respuesta del tipo "tiempo hasta evento" con censura. Se ha evaluado el método propuesto con otros dos métodos ya existentes y implementados en muchos *softwares*, como son el método Naive y el Dosage.

El método Naive no tiene en cuenta la incertidumbre: al estimar la asociación, se ajusta un modelo de regresión para datos censurados (Cox o de Weibull) asignando a cada individuo el genotipo más probable derivado de la imputación. Por el contrario, el método Dosage sí que tiene en cuenta la incertidumbre, pero de una forma "especial": también ajusta un modelo de Cox o Weibull introduciendo el SNP imputado como variable explicativa, pero no como el genotipo más probable sino como la esperanza del número de copias según las probabilidades resultantes de la imputación. A esta esperanza también se la conoce como *dosage*. Esta técnica puede producir valores esperados extraños, en el sentido que mientras que un SNP no pue-

de tomar más de 3 valores distintos (genotipos), el *dosage* es una variable continua que puede tomar cualquier valor entre 0 y 2.

Los modelos de Cox y de Weibull son los más usados para el análisis de datos longitudinales. El primero es semiparamétrico, mientras que el segundo es completamente paramétrico, siendo la regresión de Cox algo más robusta. Dado que el método propuesto necesita de la expresión explícita de la función de verosimilitud, se ha descartado adaptar el modelo de Cox. En cambio sí que se ha podido adaptar el modelo de Weibull ya que es totalmente paramétrico.

A priori, ninguno de los dos métodos existentes (ni el Naive ni el Dosage) parecen satisfactorios: el Naive porque no tiene en cuenta la incertidumbre y esto se sabe que lleva a resultados sesgados y poco potentes, y el método Dosage por incorporar en el modelo una variable continua que no equivale al SNP (categórica con 3 niveles o categorías). Otros estudios anteriores han mencionado que este segundo método (Dosage) da resultados insesgados pero no lo justifican. Por lo tanto, en este artículo se ha tratado de cuantificar en qué escenarios de incertidumbre, frecuencia alélica, magnitud de la asociación, etc., los métodos existentes son suficientemente buenos o dan resultados poco sesgados. Estos resultados se han comparado con el método propuesto, que en teoría incorpora de forma apropiada la incertidumbre. Concretamente, se ha medido el sesgo y la potencia estadística de los estimadores de la asociación del SNP imputado y el tiempo hasta el evento. También se han analizado datos reales procedentes del estudio de Framingham con aproximadamente 2,5 millones de SNPs imputados y una cohorte de más de 3.000 participantes con un seguimiento de unos 10 años de eventos cardiovasculares (variable respuesta).

En el estudio de datos simulados no se han generado datos de una Weibull (lo cual invalidaría los resultados al usar el mismo modelo para generar los datos que para ajustarlos) sino que se han generado a partir de la función empírica derivada de los datos de la cohorte de Framingham. Por último, para los métodos Naive y Dosage se ha ajustado una regresión de Cox, incluyendo el genotipo más probable y la esperanza (*dosage*) como variable explicativa, respectivamente.

Resultados

Los resultados del estudio con datos simulados han demostrado que el método propuesto no tiene sesgo, incluso en los escenarios de alto grado de incertidumbre y

gran magnitud de asociación (Hazard Ratio), mientras que el método Dosage presenta algo de sesgo cuando la incertidumbre es alta y la asociación también lo es. En el resto de escenarios (incertidumbre y asociación bajas o moderadas) estos dos métodos han dado resultados muy similares. Por el contrario, el método Naive ha resultado sesgado o muy sesgado en todos los escenarios, excepto, como era de esperar, cuando no hay asociación. Por último, los métodos Dosage y el propuesto han alcanzado una potencia estadística similar en todos los escenarios, mientras que el método Naive ha sido mucho menos potente.

Los resultados del análisis de los datos reales procedentes de la cohorte de Framingham han mostrado como ninguno de los tres métodos aporta un exceso de falsos positivos. En cuanto al tiempo requerido para el análisis de todos lo SNPs imputados analizados (más de un millón en total), el método propuesto ha sido sólo 2 ó 3 veces más lento que los otros dos. Así, pues, y desmintiendo a otros estudios, se ha demostrado como el método propuesto, aunque ajuste un modelo más complejo que los "convencionales" (regresión lineal, logística o de Cox), no es mucho más costoso. Una explicación de ello es la rapidez del método implementado para ajustar el modelo propuesto (algoritmo de N-R), donde además se han facilitado las derivadas analíticas.

5.3.3. Artículo 3: *Interaction association analysis of imputed SNPs in case control and longitudinal studies.*

Métodos

Hasta la fecha, pocos GWAS han conseguido explicar un porcentaje alto de la variabilidad genética de la enfermedad estudiada. Para solventarlo, se han estudiado otras variantes genéticas diferentes de los SNPs, como son los CNVs. Además, gracias a la imputación, se han podido incorporar más SNPs en los estudios. Aún así sólo se ha conseguido explicar una parte pequeña de la variabilidad genética en muchas enfermedades. Otra alternativa a los CNVs ha consistido en el estudio de las interacciones de SNPs, también conocido como epistasis. La hipótesis es que parte de la variabilidad no explicada reside en el exceso de riesgo que aporta una combinación de SNPs respecto a la suma de todos ellos por separado.

Existen varios modelos estadísticos para estimar la asociación de varios SNPs y la variable respuesta incluyendo las posibles interacciones: regresión lógica, árboles de regresión, *random forests*, etc. Aunque ninguno de ellos está pensado (que se sepa hasta la fecha) para incorporar incertidumbre en los SNPs imputados. Así que su aplicación queda limitada a SNPs genotipados o imputados con muy poca incertidumbre. Por otro lado, los *softwares* que pueden tratar con SNPs imputados (usando el método Naive o el Dosage) no están diseñados para los estudios de interacción con SNPs imputados.

En este trabajo se ha presentado una extensión del modelo propuesto pero ahora incluyendo interacciones de pares de SNPs imputados, tanto para estudios de caso-control como para estudios de cohorte o longitudinales (con seguimiento). Se ha comparado el método propuesto con los equivalentes a los métodos Naive y Dosage. A pesar de que este último (Dosage) no está implementado en ningún *software*, es fácil formularlo: se trata de incorporar los *dosages* de los dos SNPs y su interacción o producto como variables explicativas en una regresión de Cox (estudios de cohorte) o logística (caso control ó respuesta binaria). El método Naive consiste simplemente en asignar el genotipo más probable a cada uno de los dos SNPs imputados y proceder como si no hubiera incertidumbre.

Para evaluar el método propuesto y los dos métodos existentes (Dosage y Naive) se ha calculado el sesgo y la potencia estadística sobre datos simulados bajo distintos

escenarios variando la frecuencia alélica y el grado de incertidumbre de ambos SNPs así como la magnitud del efecto de la interacción. Además, se han analizado datos reales procedentes de la cohorte de Framingham con más de 3.000 participantes. Se ha realizado un estudio de interacción sobre aquellos SNPs que hayan alcanzado una asociación marginal con un p-valor inferior a 0,001 de forma univariada. De esta forma, se han analizado aproximadamente un millón de pares de SNPs. A este tipo de estudios de interacciones (epistasis) a lo largo de todo el genoma se les denomina GWIS.

Así como existen algunos estudios evaluando los métodos existentes (Naive, Dosage,...) para SNPs imputados en los GWAS, no hay ninguno hasta la fecha que haya analizado el efecto de la incertidumbre en la imputación en los GWIS. Sin embargo, es de vital importancia estudiar el efecto de la incertidumbre en los GWIS con SNPs imputados, ya que la incertidumbre crece exponencialmente cuando se combinan SNPs. En este trabajo se ha estudiado en qué situaciones (grado de incertidumbre, magnitud de la interaccción, ...) los métodos Naive y Dosage son válidos y cuándo el método propuesto mejora significativamente a los otros dos.

Resultados

A partir de los resultados sobre datos simulados, se ha visto como el método propuesto es insesgado en todos los escenarios, incluso en situaciones extremas donde el grado de incertidumbre y la magnitud de la interacción son elevados. En cambio, el método Dosage ha presentado un sesgo negativo o positivo dependiendo de la frecuencia alélica, mientras que el método Naive ha sido sesgado en todos los escenarios con algún grado de incertidumbre. En cuanto a la potencia estadística, el método propuesto y el Dosage se han comportado de forma similar y en algunos escenarios el primero ha sido mejor, sobretodo cuando la magnitud de la incertidumbre y el efecto de la interacción son elevados. Finalmente, el método *Naive* ha conseguido bastante menos potencia estadística que los otros dos.

Del estudio con datos reales, los tres métodos han dado resultados similares, y ninguna interacción ha alcanzado la significación estadística. Por otro lado, el tiempo requerido para analizar los datos reales con el método propuesto no ha sido mucho mayor que para los otros dos, los cuales utilizan modelos "convencionales" (regresión de Cox). Así pues, se ha podido concluir que es indistinto, en tiempo computacional, utilizar los tres métodos (inclusive el modelo propuesto) para analizar este tipo de

estudios (GWIS).

Cabe destacar que el punto de corte para considerar que hay demasiada incertidumbre en los estudios de interacción (GWIS) no puede ser el mismo que para los estudios sin interacciones (GWAS), sino que hay que ser más restrictivo. Esta estrategia conlleva descartar muchos más SNPs a priori. Por lo tanto, es muy recomendable usar alguna estrategia que tenga en cuenta la incertidumbre, ya sea el método Dosage o el propuesto, siendo este último especialmente recomendable cuando la incertidumbre es mayor.

5.3.4. Artículo 4: *CNVassoc: Association analysis of CNV data using R.*

El objetivo de este artículo ha sido implementar una herramienta informática que permita ajustar los modelos de asociación entre CNVs y la variable respuesta para estudios de caso-control, cohorte con seguimiento, de respuesta continua o conteos, usando el modelo propuesto presentado en los artículos anteriores de la tesis, haciendo posible ajustar por covariables, modificar el modelo de herencia a aditivo, dominante, recesivo o co-dominante, etc. Aunque el modelo estadístico se pensó inicialmente para analizar CNVs, el programa permite analizar también SNPs imputados.

A parte de las funciones propias para estimar el modelo de asociación entre el CNV o el SNP imputado y la variable respuesta, el *software* permite utilizar otras funciones útiles para inferir el número de copias o para graficar la intensidad de la señal de la sonda obtenida durante el genotipado de los CNVs, etc. A fin de incrementar su rapidez en los cálculos en los estudios GWAS o con múltiples CNVs, se han incorporado funciones que realizan los cálculos en paralelo y en `C++`.

Todas las instrucciones que configuran el programa implementado se han estructurado conformando un paquete dentro del software `R`. Ello ha permitido una mejor documentación de todas las funciones y una optimización organizando el código en objetos, clases y métodos. El paquete desarrollado se llama `CNVassoc` y está disponible en el repositorio CRAN de `R` `http://cran.rstudio.com/`. Para ilustrar su funcionamiento, se ha escrito un extenso manual con ejemplos en que se analizan datos reales disponibles en el mismo paquete.

Los motivos para implementar la herramienta en `R` han sido varios: `R` es uno de los programas estadísticos más usado cuya comunidad de usuarios es cada vez mayor, es gratuito y de código abierto, y muy flexible. A diferencia de otros programas existentes para realizar GWAS cómo PLINK, el código implementado en `R` se puede modificar, así como los "inputs" (ficheros de datos) y los "outputs" (resultados), etc.

Otro objetivo de este artículo ha sido comparar la eficiencia y utilidad del paquete `CNVassoc` con otro existente para el análisis de asociación con CNVs llamado `CNVtools` implementado también en `R`. La metodología con la que se basa `CNVtools` es más compleja que `CNVassoc` dado que realiza la inferencia del número de copias

y la estimación de los parámetros de asociación en un solo paso. Aunque desde el punto de vista teórico, esta estrategia sea más sólida, la computación es mucho más lenta y en situaciones de moderada o gran incertidumbre se ha visto que el modelo no converge. Los resultados del análisis sobre datos reales disponibles en `CNVtools` han sido muy parecidos para ambos paquetes, aunque `CNVassoc` ha demostrado ser mucho más robusto y rápido.

APÉNDICE

6.1. Copia de las artículos en revisión (artículo 3)

Interaction association analysis of imputed SNPs in case control and longitudinal studies

Isaac Subirana[1,2,3] and Juan R González [4,1,5*]

[1]CIBER Epidemiology and Public Health (CIBERESP), Spain
[2]IMIM, Parc de Salut Mar, Spain
[3]Department of Statistics, University of Barcelona, Spain
[4]Center for Research in Environmental Epidemiology (CREAL), Spain
[5]Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Spain

Email: Isaac Subirana - isubirana@imim.es; Juan R González*- jrgonzalez@creal.cat;

*Corresponding author

Abstract

A new method is described to assess the interactions of imputed SNPs (single nucleotide polymorphisms) in case-control and cohort studies, properly incorporating SNP imputation uncertainty in the likelihood model. Using simulation studies and analysis of real data obtained from the Framingham study cohort, we compare the performance of this new method to the DOSAGE and NAIVE (also known as Best-Guess) methods, developed in the context of single SNP and extended to SNP-by-SNP interaction. The results show that only our new method is unbiased under all examined scenarios regarding allele frequencies, imputation uncertainty degree and interaction effect size. In addition, our method achieves at least as much power as the other two, and exceeds their statistical power in certain cohort analysis situations. This method has been implemented in C-code and integrated into R software, and is fast enough to perform Genome Wide Interaction Studies (GWIS) with hundreds of thousand of interactions.

Introduction

Genome-wide association studies (GWAS) have become very popular in the last decade as the power of genotyping technology has increased. It is now feasible to conduct studies with a million Single Nucleotide Polymorphisms (SNP), spread throughout the whole genome. However, to date, GWAS have identified only a small proportion of heritability in most of the common complex diseases [1]. For example, it has been

estimated that genetic factors explain between 35% and 60% of thrombosis cases, but all identified SNPs explain only 5% of the phenotype variability [2].

In order to overcome these difficulties, genetic variants other than SNPs, such as copy number variants, have been studied with some success [3]. Another approach to explain missing heritability consists in analyzing interactions between SNPs (epistasis) where moderate and large interaction effects between SNPs have been found [4]. This contrasts with the small effects usually estimated in GWAS, with odds ratios ranging from 1.1 to 1.5 in common or complex diseases (e.g., [5–10])). Therefore, in the GWAS context, studies of SNP interaction (known as GWIS) have received much attention in recent years.

Currently, GWAS/GWIS incorporate information about non-genotyped SNPs by incorporating the information on the surrounding genotyped SNPs and Hap Map platform (e.g., imputed SNPs). This has been made possible due to major efforts to develop bioinformatics tools for imputing SNPs (see a description of existing algorithms in [11]). Estimates of genotypes provided by existing methods must be considered to contain uncertainty, which must be included in association analysis in order to avoid biased and/or underpowered estimates [12–14]. Therefore, statistical analysis of GWIS with imputed SNPs should also be performed by incorporating such uncertainty in the models. There are several strategies to address this issue in the context of single association analysis (e.g., GWAS). The first places the most probable genotype for each individual as if it were the genotype (NAIVE). The second places the expected number of alleles for each individual according to the imputed genotype probabilities (DOSAGE). Finally, the third approach maximizes the proper likelihood by using a latent class (LC) model. For case-control studies, the DOSAGE approach is implemented in SNPTEST [11], PLINK [15] and ProbABEL [16], while LC is available at `CNVassoc R` package [17]. Version 2 of SNPTEST software also deals with quantitative traits, and `CNVassoc` is capable of analyzing quantitative [13] and survival data [18].

Several authors have analyzed the behaviour of these three approaches in GWAS settings for both binary and quantitative variables. The main conclusion is that for small effects, DOSAGE and LC methods behave correctly in all situations, while NAIVE is biased and underpowered [13, 14]. For moderate and large effects, LC outperforms the DOSAGE approach, and the advantage increases with greater uncertainty [13]. Similar results are obtained when analyzing survival data [18]. Nonetheless, no studies have explored how uncertainty can affect the results of SNP-by-SNP interaction analysis in the case of imputed data.

Theoretically, interaction effects should be much greater than those observed when analyzing a single SNP. This would imply that LC would provide better estimates than the DOSAGE approach. In addition, uncertainty is much higher in the case of analyzing the interaction between two imputed genotypes than in the analysis of a single SNP because two sources of uncertainty are combined. This would also favour using the LC approach in order to increase the power to detect statistically significant interactions. A large number of bioinformatics tools have been described for analyzing GWIS (see an exhaustive list in [19]). However, the vast majority of them are limited to case-control studies and none of them supports imputed SNPs.

In this paper we aim to address the existing limitations when analyzing SNP-by-SNP interaction with imputed SNPs, from both a theoretical and practical point of view. First, we formulate the theory behind the different strategies (NAIVE, DOSAGE and LC) to assess interaction of imputed SNPs in both case-control and longitudinal studies. Then, using exhaustive simulation studies we investigate the behaviour of these three approaches in terms of bias and power. The simulations model a wide range of the exhaustive possibility of scenarios with varying allele frequencies, interaction effects and degree of uncertainty in imputed SNPs. Simulation results are complemented by GWIS using $\sim$ 2.5M of imputed SNPs from a Framingham cohort data set collected by a longitudinal study on time-to-coronary event (`http://www.framingham.com/heart/`). Finally, `R` functions implementing the proposed method have been included in the `CNVassoc` package.

Methods

Model

In this section, we provide the formal association model for imputed SNPs interactions and a trait adjusting for possible covariates, such as age, sex, etc., or principal components to take into account in population stratification. We develop the model's likelihood for a case-control study (binary trait) or cohort study (Weibull distributed trait). Finally, we discuss theoretical aspects of this model, and compare different strategies to approach the parameter estimation.

Likelihood

When uncertainty is present in imputed SNPs, the likelihood function takes the form (applying the Bayes' theorem):

$$L(\mathbf{Y};\boldsymbol{\Theta}) = \prod_{i=1}^{N} \sum_{k=0}^{2} \sum_{l=0}^{2} \text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l, \boldsymbol{C}_i; \boldsymbol{\Theta})\, \text{Prob}(\text{SNP}_1 = k, \text{SNP}_2 = l|i) \tag{1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is the vector of observed values of the response variable for all sample subjects, $\text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta})$ is the probability or density function for continuous or discrete response, respectively, given the number of alleles for SNP_1 and SNP_2, $\boldsymbol{\Theta}$ are the model parameters (the odds ratio for a logistic regression or hazard ratio for a time-to-event response, etc.), $\boldsymbol{C}_i$ is the vector of covariates (e.g. sex, age, etc.) and $\text{Prob}(\text{SNP}_1 = k, \text{SNP}_2 = l|i)$ is the probability of having k and l risk alleles in the first and second SNP, respectively, for the i-th individual. If independence in genotype imputation is assumed, this last probability can be decomposed as the product of genotype imputation of each SNP, which can be seen as the outer product of the 3 vector probabilities:

$$\text{Prob}(\text{SNP}_1 = k, \text{SNP}_2 = l|i) = \text{Prob}(\text{SNP}_1 = k|i)\, \text{Prob}(\text{SNP}_2 = l|i) \tag{2}$$

Note that the likelihood function (1) has no standard form because it contains a double summation for each individual. Therefore, no "standard" regression methods can be used to maximize it, i.e. to find the Maximum-Likelihood (ML) estimates of parameters ($\hat{\boldsymbol{\Theta}}$).

$\text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta})$ could be estimated in the same model. This approach is also known as one-step algorithms. Although this could be a good strategy from a theoretical point of view [20], calculation of these algorithms is not feasible when SNP imputation involves hundreds of SNPs in Linkage Disequilibrium (LD) from the same sample or/and from an external Hap Map reference sample.

Another much more practical approach to maximize (1) consists of taking the SNP probabilities previously estimated from any imputation algorithm such as IMPUTE [21, 22] or MACH [23], instead of estimating them directly from the likelihood equation (1). However, imputation algorithms do not provide pairwise joint imputation probability, and therefore independence in SNP imputation must be assumed. Although this assumption is difficult to verify, it does not seem unrealistic. This strategy is known as a two-step algorithm, which first estimates the probability of having each genotype (aa, aA or AA) for each individual and each imputed SNP (using any imputation software), and then uses these estimated or imputed

probabilities to estimate the association model, i.e., to maximize (1).

Once $\text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta})$ is known or estimated by any imputation algorithm, maximization of (1) is feasible and not much more computationally demanding than standard regression (e.g. logistic or Cox regression). To search for ML estimates we took advantage of the Newton-Raphson (NR) procedure, which is known to be very fast in maximizing non-linear functions, i.e., converging to the maximum in just a few steps. Also, using NR can obtain not only a point estimate of parameters but also their standard errors and thus their confidence intervals. To speed up the process, we derived analytic first and second derivatives instead of simply plugging numerical ones into the NR procedure, which multiplied the speed by more than 10. A very detailed description of the likelihood functions for a case-control and cohort study, as well as their first and second derivatives required to performed the NR procedure in assessing interaction of imputed SNPs, is shown in Additional file 1.

Case-control studies

When analyzing data from a case-control study, it is common to fit a logistic regression model where the response is coded as 0 for controls and 1 for cases, and SNPs and possible other variables (covariates, adjusting or confounding variables) are taken as predictors. When SNPs are not observed but imputed, the likelihood function (1) is taken, where

$$\text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta}) = \frac{e^{y_i \eta_{ikl}}}{1 + e^{\eta_{ikl}}} \tag{3}$$

and the linear predictor (η_{ikl}) is:

$$\eta_{ikl} = \beta_0 + \beta_1 k + \beta_2 l + \beta_{12} kl + \sum_{j=1}^{Q} \gamma_j C_{ij} \tag{4}$$

where k and l are the number of risk alleles (0, 1 or 2) for the first and second SNPs, respectively, and C_{ij} is the value of the j-th covariate (predictor different from the SNPs) for the i-th sample individual.

Therefore the parameter vector, $\boldsymbol{\Theta}$, consists of the constant, (β_0), response probability when all predictors are equal to zero; main effect slope (β_1 and β_2); log-Odds Ratio of incrementing one allele of the first SNP and the second SNP, respectively; the interaction (β_{12}) term, log-Odds Ratio of incrementing one extra risk allele for one of the two SNPs; and covariate coefficients (γ_j, $j = 1, \ldots, Q$), log-Odds Ratio for each possible covariate.

According to the way that the linear predictor is defined, an additive effect for both SNPs as well as for the interaction term is supposed. Although several combinations of other types of genetic effects (e.g. dominant, recessive or codominant) for each main effect and interaction term can be considered, we have considered only the additive effect, both for simplicity and because it is the most used GWAS or GWIS model of inheritance.

Cohort studies

Cohort studies may attempt to predict not only whether or not an individual had an event, but also when it occurs. Therefore, response consists of two variables, Y and δ, where Y is the observed time and δ is the censored variable indicator, coded as 0 if the event does not occur during the follow-up time and 1 otherwise.

The most commonly used models to analyze this type of data are Proportional Hazard Cox regression and Weibull regression. The first is a semiparametric model, the second is fully parametric. Although the Cox regression does not make any assumption about response distribution and therefore is more flexible than Weibull, the latter fits well on the vast majority of real data. Moreover, it is not feasible to adapt the likelihood function for imputed SNPs (1) to a model like Cox regression that is not fully parametric. We implemented the likelihood function when assuming a Weibull distributed response (Weibull regression) to imputed SNPs. In this case, the above probability takes the form:

$$\text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta}) = \begin{cases} e^{-\lambda_{ikl} Y_i^{\phi}} & \text{if } \delta_i = 0 \\ \lambda_{ikl}\phi Y_i^{\phi-1} e^{-\lambda_{ilk} Y_i^{\phi}} & \text{if } \delta_i = 1 \end{cases} \tag{5}$$

where $\lambda_{ikl} = exp(\eta_{ikl})$ and η_{ikl} is the linear predictor defined as for case-control studies, except that now *beta* and γ coefficients are interpreted as log-Hazard Ratios instead of log-Odds Ratio. The scale parameter (ϕ) is added to this calculation.

Estimation strategies

The main goal of this paper is to compare different strategies to estimate the interaction effect (β_{12}). More specifically, we want to compare a method we propose, following [13] but extended to incorporate SNP-by-SNP interaction (Latent Class), to two other widely used methods, DOSAGE and NAIVE.

All three strategies are two-step approaches, as they take previously imputed probabilities obtained by any

imputation software and fit the association model (logistic regression, Weibull regression, etc. regression). The differences between them exist mostly in the way they incorporate these probabilities into the association likelihood model.

- **Latent Class (LC):** The LC strategy uses the proper likelihood of the model (1), and estimates the actual parameters by Maximum Likelihood. Therefore, by the maximum likelihood theorem, estimators are unbiased and achieve the minimum variance, at least asymptotically. This method uses the NR procedure to obtain the ML estimates.

- **NAIVE:** The NAIVE strategy, also known as Best-guess, consists in fitting a classical model (linear, logistic, Weibull, etc., depending on the response distribution) to the data, introducing the SNPs as if they were genotyped or observed; for each individual, the most probable genotype for each SNP is placed. It is known that this strategy does not capture the possible uncertainty in SNP imputation, since it is not the same to have 0.95, 0.01 and 0.04 chance of having 0, 1 or 2 copies, compared to 0.5, 0.3 and 0.2. In both cases, the most probable genotype is 0 risk alleles but it is clear that in the second case there is much more uncertainty. The consequence is that the effect estimation is biased and underpowered.

- **Probabilities as covariates (DOSAGE)** The DOSAGE approach is similar to the NAIVE approach in the sense that both fit classical models, but the first one introduces the expected number of alleles instead of the most probable genotype. Thus, DOSAGE approach captures more information about uncertainty.

Since DOSAGE and NAIVE methods fit classical models, they are computationally less demanding and standard software is available to fit them. However, there is no available genetic software that supports SNP-by-SNP interaction for imputed SNPs using the DOSAGE approach. Moreover, the LC approach, which maximizes the proper likelihood, has not yet been implemented in any software.

Another key difference between fitting a complex likelihood (LC strategy) or a classical model (DOSAGE or NAIVE strategies) is that the first requires more sample size to converge. Therefore, for small samples and/or rare SNPs, the LC strategy may have some difficulty obtaining satisfactory estimates, while the other two strategies (NAIVE or DOSAGE) may give better estimations. By contrast, for big enough samples or less rare SNPs, the LC strategy may give better results in terms of accuracy and power.

Simulation

Data simulation

We carried out two simulation studies to examine the behaviour of interaction effect (β_{12}) estimation when using each of the three strategies described in previous section. The first study investigates their behaviour in the context of case-control studies, while the second one focuses on longitudinal data. Data simulation was performed using data from the Framingham Study (`http://www.framingham.com/heart/`) that contain more than 2 million of imputed SNPs. The data were simulated as follows:

1) **Imputed SNPs:** Imputed SNPs were randomly selected from the Framingham heart cohort study (`http://www.framingham.com/heart/`), a sample with more than 8,477 individuals, about a million genotyped SNPs and approximately 2.5 million imputed SNPs. Different Minor Allele Frequencies (MAF) and different degrees of imputation uncertainty were considered. The uncertainty measure used was R^2, defined as the correlation between actual genotype and most probable genotype. This definition of R^2 is incorporated by imputation software MACH, which was used to impute SNPs in the Framingham cohort. R^2 ranges from 0 (no information, huge uncertainty) to 1 (complete information, no uncertainty). We selected 100 SNPs from each MAF combination [0.05, 0.15), [0.25, 0.35) and [0.45, 0.50) and R^2 in [0.05, 0.15), [0.25, 0.35), [0.45, 0.55), [0.65, 0.75) and [0.85, 0.95), covering from rare to common SNPs and from very low to very high uncertainty. Table 1 shows the total number of imputed SNPs from the Framingham cohort in each of these combination of R^2 by MAF bins.

2) **Response simulation:** Response was simulated by genotype status, i.e., the number of risk alleles. Since imputed Framingham genotype data do not include any genotyped SNPs, they were generated randomly: from each individual and each imputed SNP, one of the 3 possible genotypes was sampled from the imputed probabilities. Therefore, we proceeded differently from [14]. Of course, our strategy in obtaining the genotype assumes that the imputed probabilities are correct. We adopted this strategy for two reasons: (i) it saves a lot of time in avoiding the need to perform SNP imputation again (Framingham data contains already imputed SNPs), and (ii) it removes the possible effect of imputation bias in assessing the association test, which is not the aim of our study. This step is required to simulate the response in the next step, which is generated given the genotype of simulated SNPs.

   After obtaining the genotype status, response was generated for the case-control study on one hand and

for cohort studies on the other.

For cohort study simulation, we simulate the response assuming no specific distribution. Only proportional hazard was supposed:

$$F(t) = 1 - S(t)^{\exp(\beta_1 \,\mathrm{SNP}_1 + \beta_2 \,\mathrm{SNP}_2 + \beta_{12} \,\mathrm{SNP}_1 \,\mathrm{SNP}_2)}$$

The baseline survival function ($S(t)$) was generated from the empirical observed time-to-coronary event from Framingham cohort. Using the previous equation, observed incidence strongly depends on risk allele frequencies as well as main and interaction effects. To avoid this we used the following equation instead:

$$F(t) = 1 - S(t)^{\exp(\beta_1 \,\mathrm{SNP}_1 + \beta_2 \,\mathrm{SNP}_2 + \beta_{12} \,\mathrm{SNP}_1 \,\mathrm{SNP}_2 - \mu - \sigma^2/2)}$$

where μ and σ^2 is the expectancy and the variance of $\beta_1 \,\mathrm{SNP}_1 + \beta_2 \,\mathrm{SNP}_2 + \beta_{12} \,\mathrm{SNP}_1 \,\mathrm{SNP}_2$, respectively. Using this equation guarantees obtaining approximately the same incidence for all simulated scenarios (8%). To simulate the response from the Framingham cohort, two Kaplan-Meier curves were performed, one taking the observed events and the other the censored events. From the first one, time-to-event values, T, were generated, and censoring times, C, from the second one. The observed times, Y, have been defined as $Y = min\{C, T\}$. Finally, if $C > T$, the value was considered censored, and non-censored otherwise (i.e. $C \leq T$). For all strategies, the rate of non-censored events was similar to the observed coronary events rate in the Framingham cohort (i.e., 8% approximately).

For case-control study simulation, we simulated a binary variable given the SNPs genotypes with probability:

$$\mathrm{Prob}(Y = 1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 \,\mathrm{SNP}_1 + \beta_2 \,\mathrm{SNP}_2 + \beta_{12} \,\mathrm{SNP}_1 \,\mathrm{SNP}_2 - \mu)}$$

where μ is defined the same as for the cohort study. The constant β_0 is set to zero, to mimic a case-control study with as many cases as controls; subtracting μ guarantees approximately 50% cases and controls independently of the simulated scenario (i.e. risk allele frequency, uncertainty degree and interaction effect.

3) **Estimation:** For each simulated data element (response, non-observed genotype, and imputed SNP probabilities for the two SNPs), 3 models were adjusted corresponding to LC, DOSAGE and NAIVE strategies. For the case-control simulation study, a logistic model was fitted, taking the most probable genotype or expected number of alleles as predictors for NAIVE and DOSAGE, respectively, while for LC the proper likelihood function for case-control studies described in the Simulation section was maximized.

For the cohort simulation study and for NAIVE and DOSAGE, we proceeded in the same way as for case-control, but fitting a proportional hazard Cox model. However, for the LC strategy, it is not feasible to fit a Cox model and a fully parametric model had to be used instead; more concretely, a Weibull regression described in the Model section.

Different scenarios were simulated, varying the minor allele frequency (MAF), degree of uncertainty (R^2) and interaction effect (β_{12}). For the interaction effect, we simulated scenarios with no effect ($\beta_{12} = 0$), which is equivalent to a Hazard Ratio, or Odds Ratio, of 1 for cohort or case-control simulation studies, respectively, to moderate, high or very high effect (i.e. HR/OR $\in$ 1.5, 2, 2.5, 3). This results in 450 different scenarios, taking into account different MAF and R^2 for both SNPs and different interaction effects. Finally, for both case-control and cohort simulation studies, SNPs main effects (β_1 and β_2) were fixed to zero. We repeated each step 1,000 times for each scenario.

In each simulation and model, the following measures were computed: 1) Bias: difference between the expected and the true interaction OR or HR; 2) Power: Power of detecting an associated SNP; and 3) Coverage: Probability that the 95% confidence interval includes the true interaction OR or HR.

Simulation results

The results obtained from all 450 different simulated scenarios were tabulated in Supplementary Table 1 and 2 for cohort and case-control simulation studies, respectively. Both tables contain bias, power, and coverage as well as failure rate obtained in the simulated study for all 450 scenarios.

Due to the large number of simulated scenarios, we summarized them in figures split into different panels according to MAF and R^2, where we illustrated only the scenarios with the same MAF and R^2 in both SNPs, for simplicity (Figures 1 to 3 for cohort and Supplementary Figures 1 to 3 for case-control simulation studies).

Cohort simulation study

The LC strategy presents no bias, regardless of the amount of uncertainty and minor allele frequency. The DOSAGE strategy begins to have bias when the effect (HR) is moderate or large ($HR \geq 2$), especially when uncertainty and MAF are large. For example, when MAF is approximately 50% and R^2 <70%, DOSAGE strategy estimates are underestimated. We also observed that DOSAGE strategy overestimates the true

HR when MAF is low (10% approximately), and has a smaller bias for intermediate MAF (around 30%). Finally, in almost all scenarios, the NAIVE strategy largely underestimates the true HR, except in the scenarios with very little uncertainty (both SNPs with R^2 >85%).

The LC and DOSAGE strategies are much more powerful than NAIVE for all scenarios with a certain amount of uncertainty and LC achieves at least as much power as DOSAGE in all scenarios. It is considerably more powerful when there is more uncertainty and MAF is high. For example, when MAF is approximately 50% for both SNPs and R^2 is around 50% for both SNPs, LC has a power of almost 80% while DOSAGE does not reach 40% for a HR=2.5.

The LC strategy presents an observed coverage almost identical to the nominal one (i.e., 95%) in all scenarios. By contrast, DOSAGE strategy coverage dips below 90% when uncertainty and HR increase; for example, when MAF is approximately 50% and R^2 is around 50% for both SNPs and HR=2.5, observed coverage for DOSAGE strategy is below 70% while for LC it is 95%. NAIVE strategy had a very rapid decrease in observed coverage when HR was increased for all MAF and R^2 scenarios, even when uncertainty was not high.

Regarding failure in model convergence, the DOSAGE strategy failed less; NAIVE and LC failure rates were similar. However, in most of the scenarios, all three methods converged in fitting all the simulated data. The most critical scenarios have very high uncertainty and small MAF. In the most extreme scenario, i.e., R^2 around 10% and MAF around 10% for both SNPs, both LC and NAIVE strategies failed 30% and even 70% of the times depending on HR, while DOSAGE could fail up to 20% of the times. However, when LC could achieve results (converged), this has been shown to be less biased and more powerful than the other two strategies.

Finally, the results show no type I error inflation or excess of false positives. Although it is not possible to demonstrate that observed power approximates the GWAS significance level due to the limited availability of simulated data for each scenario (i.e., 1,000), it was very low (< 0.001) for all simulated scenarios with no interaction effect (OR=1) for all three strategies. Also, observed coverage approximates 95% and there is no or very little bias for all strategies with no interaction effect scenarios, except for the most extreme scenario (with very high uncertainty and low MAF for both SNPs) where NAIVE and LC strategies slightly overestimate HR.

Case control simulation study

Supplementary Table 2 and Supplementary Figures 1 to 3 show that interaction estimation properties for all three strategies were similar in terms of bias and coverage with respect to the cohort simulation study results: LC is not biased and achieves the desired 95% coverage, while DOSAGE is biased when effect is moderate or large (OR bigger or equal to 2) and uncertainty increases, and Naive method largely underestimates the true OR in all scenarios.

However, LC is not more powerful than DOSAGE, in contrast from what was seen in the cohort simulation study, and both strategies achieve similar power. Finally, as in the cohort simulation study, the NAIVE strategy has much less power than LC or DOSAGE strategies.

We observed results very similar to those obtained in the cohort regarding scenarios with no interaction effect. Therefore, no type I error inflation for any of the three methods is observed when simulating a case control study.

Analysis of real data

Aside from performing a simulation study taking imputed SNPs from the Framingham study data, we also took real phenotype data from the same study. We obtained access to phenotype and genotype (imputed SNPs) data under the Framingham Share initiative via the Database of Genotypes and Phenotypes (dbGaP, ncbi.nlm.nih.gov/dbgap; Project number 1534). From all imputed SNPs existing in the database, we discarded all the non-common ones (minor allele frequency $\leq$10%) and the ones imputed with almost no uncertainty ($R^2 \geq 99\%$) or too much uncertainty ($R^2 \leq 10\%$).

Before performing the interaction analysis, we conducted a single-SNP association analysis of each imputed SNP and time-to-coronary event (response), adjusting by age, sex and the first five principal components to take into account an excess of familial or ethnic data structure. To do so, the three strategies (Naive, DOSAGE and LC) were fitted on the data similarly to the description in the Simulation section but assuming an additive genetic effect for a single SNP.

The SNPs achieving marginal significance with a p-value$<$0.001 in the univariate analysis were further analyzed in the interaction analysis, similarly to the strategy adopted by [24]. The idea is that those SNPs

interacting with others will have a non-null side effect, in the sense that some effect should be observed when performing association with one-by-one response. The significance level of 0.001 was chosen to make the interaction study feasible, i.e., expecting about 1,000 SNPs to be selected for interaction analysis, which translates to about a million of interactions.

Finally, we performed an association test of SNP-by-SNP interaction, adjusting by the same covariates used in the single SNP association (i.e., sex, age and five principal components). In all analyses, p-values were computed after correcting by genomic inflation factor λ to address an excess of familial or ethnic data structure of the data not captured by the first five principal components.

Results

We analyzed 3,007 individuals from the offspring cohort with complete data available, i.e., imputed SNPs, time-to-coronary event, age, sex and five principal components. Among the 1,021,307 imputed SNPs tested in the univariate association, 1,334 achieved the significance level of 0.001 in at least one of the three strategies (LC, DOSAGE or NAIVE). Therefore, 889,111 pairs of SNPs were further tested in the interaction analysis.

None of the three strategies gave any significant result after Bonferroni correction (α=5.6e-08). The most significant interaction achieved a p-value of 1.95e-06 after correcting for the genome inflation factor. According to the QQ-plot of p-values (Figure 4), there seemed to be no type I error inflation. The estimated genomic inflation factor, λ, was very close to one for all strategies (1.054, 1.044 and 1.056 for NAIVE, DOSAGE and LC, respectively), suggesting that adjusting for the principal components was sufficient.

Analysis of 889,111 pairs of imputed SNPs in 3,007 individuals and adjusting for age, sex and five genetic principal components took 11.0, 11.3 and 37.6 hours for NAIVE, DOSABE and LC methods, respectively, using a Linux platform x86_64-redhat-linux-gnu (64-bit) CPU with 2GHz and 16G of RAM Memory, and `R` software version 2.15.1. Therefore, using a multicore with 10 processors, for example, it may take less than 4 hours to carry out a similar GWIS analysis.

Finally, the failure rate (i.e., the number of imputed SNPs for which models have not converged or have achieved a non-reliable estimation) was very low for all 3 methods: less than 0.3%.

Discussion

This is the first study analyzing the properties of existing strategies in terms of bias and power to perform interactions between pairs of imputed SNPs. We have assessed how uncertainty affects the efficiency of these estimates.

According to the simulation study results, we found that the NAIVE approach is biased and underpowered in all scenarios (MAF, uncertainty and effect), while both DOSAGE and LC methods behave satisfactorily. It is important to note that some differences exist between these last two, depending on the type of study performed, case-control or longitudinal. In a case-control study, LC and DOSAGE have very similar power in all scenarios but DOSAGE gives underestimated results when the interaction effect is high, while in longitudinal studies DOSAGE overestimates results when MAF is low, and LC achieves more power when the interaction effect is moderate to large. Finally, we have observed that LC is accurate and has no bias in any scenarios for both case-control and longitudinal studies.

In many situations, it is crucial to achieve good power but also non-biased estimates. For example, (i) meta-analyses using different platforms, (ii) building prediction function of a disease where SNP-by-SNP interaction is included, or (iii) the measurement of the phenotype variability explained by genetic factors, including interactions. In the first example, we can conclude a false heterogeneity between studies and obtaining underpowered results, in the second the function may provide biased predictions and in the third we can conclude that we explain less heritability than what exists.

When analyzing real data, the three approaches show no inflation of false positives and give results concordant to other studies of coronary heart disease in case-control studies [24]. Therefore, any of them can be used without returning an excess of false positives.

Finally, our proposed method (LC) did not take much more time than the two standard approaches (DOSAGE and NAIVE), making the analysis of one million interactions with LC method doable in a standard PC.

Conclusions

The NAIVE strategy is not suggested for assessing interaction of pairs of imputed SNPs in case-control or longitudinal studies because it is biased and underpowered. Both LC and DOSAGE perform very well in almost all scenarios, with LC being the best choice when the interaction effect is bigger. Our proposed method, LC, is the only one with no bias in any scenario and its use is also feasible in a GWIS because it consumes similar time to NAIVE, DOSAGE or any other standard regression model.

Author's contributions

IS wrote the manuscript, implemented the R code and algorithms and analyzed the data. JRG designed and supervised the study and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The study was supported by a grants from Spanish Ministry of Science (MTM2011-26515 and MTM2008-02457). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278.

Authors want to thank Elaine Lilly for her contribution in the English revision of the manuscript, and Mikel Esnaola for his great contribution in translating the algorithms code syntax to C language.

References

1. Maher B: **Personal genomes: The case of the missing heritability**. *Nature.* 2008, **456**(7218):18–21.

2. Greliche N, Germain M, Lambert J, Cohen W, Bertrand M, Dupuis A, Letenneur L, Lathrop M, Amouyel P, Morange P, Trégouët D: **A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis**. *BMC Med Genet* 2013, **14**.

3. Jarick I, Vogel C, Scherag S, Schfer H, Hebebrand J, Hinney A, Scherag A: **Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis**. *Hum Mol Genet.* 2011, **20**(4):840–52.

4. Kirino Y, Bertsias G, Ishigatsubo Y, Mizuki N, Tugal-Tutkun I, Seyahi E, Ozyazgan Y, Sacli F, Erer B, Inoko H, Emrence Z, Cakar A, Abaci N, Ustek D, Satorius C, Ueda A, Takeno M, Kim Y, Wood G, Ombrello M, Meguro A, Gül A, Remmers E: **Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1**. *Nat Genet.* 2013, **45**(2):202–207.

5. Edwards T, Shrubsole M, Cai Q, Li G, Dai Q, Rex D, Ulbright T, Fu Z, Delahanty R, Murff H, Smalley W, Ness R, Zheng W: **Genome-wide association study identifies possible genetic risk factors for colorectal adenomas**. *Cancer Epidemiol Biomarkers Prev.* 2013, **22**(7):1219–1226.

6. Chung C, Kanetsky P, Wang Z, Hildebrandt M, Koster R, Skotheim R, Kratz C, Turnbull C, Cortessis V, Bakken A, Bishop D, Cook M, Erickson R, Fosså S, Jacobs K, Korde L, Kraggerud S, Lothe R, Loud J, Rahman N, Skinner E, Thomas D, Wu X, Yeager M, Schumacher F, Greene M, Schwartz S, McGlynn K, Chanock S, Nathanson K: **Meta-analysis identifies four new loci associated with testicular germ cell tumor**. *Nat Genet.* 2013, **45**(6):680–685.

7. Tanaka T, Ngwa J, van F Rooij, Zillikens M, Wojczynski M, Frazier-Wood A, Houston D, Kanoni S, Lemaitre R, Luan J, Mikkilä V, Renstrom F, Sonestedt E, Zhao J, Chu A, Qi L, Chasman D, de OM Oliveira, Dhurandhar E, Feitosa M, Johansson I, Khaw K, Lohman K, Manichaikul A, McKeown N, Mozaffarian D, Singleton A, Stirrups K, Viikari J, Ye Z, Bandinelli S, Barroso I, Deloukas P, Forouhi N, Hofman A, Liu Y, Lyytikäinen L, North K, Dimitriou M, Hallmans G, Kähönen M, Langenberg C, Ordovas J, Uitterlinden A, Hu F, Kalafati I, Raitakari O, Franco O, Johnson A, Emilsson V, Schrack J, Semba R, Siscovick D, Arnett D, Borecki I, Franks P, Kritchevsky S, Lehtimäki T, Loos R, Orho-Melander M, Rotter J, Wareham N, Witteman J, Ferrucci L, Dedoussis G, Cupples L, Nettleton J:

Genome-wide meta-analysis of observational studies shows common genetic variants associated with macronutrient intake. *Am J Clin Nutr.* 2013, **97**(6):1395–1402.

8. Lövkvist H, Sjögren M, Höglund P, Engström G, Jern C, Olsson S, Smith J, Hedblad B, Andsberg G, Delavaran H, Jood K, Kristoffersson U, Norrving B, Melander O, Lindgren A: **Are 25 SNPs from the CARDIoGRAM study associated with ischaemic stroke?** *Eur J Neurol.* 2013, **20**(9):1284–1291.

9. Trouw L, Daha N, Kurreeman F, Böhringer S, Goulielmos G, Westra H, Zhernakova A, Franke L, Stahl E, Levarht E, Stoeken-Rijsbergen G, Verduijn W, Roos A, Li Y, Houwing-Duistermaat J, Huizinga T, Toes R: **Genetic variants in the region of the C1q genes are associated with rheumatoid arthritis**. *Clin Exp Immunol.* 2013, **173**:76–83.

10. Yang Z, Shen J, Cao Z, Wang B: **Association between a novel polymorphism (rs2046210) of the 6q25.1 locus and breast cancer risk**. *Breast Cancer Res Treat.* 2013, **139**:267–275.

11. Marchini J, Howie B: **Genotype imputation for genome-wide association studies**. *Nat Rev Genet.* 2010, **11**(7):499–511.

12. Davidov O, Faraggi D, Reiser B: **Misclassification in Logistic Regression with Discrete Covariates**. *Biometrical Journal* 2003, **45**(5):541–553.

13. González JR, Subirana I, Escaramis G, Peraza S, Caceres A, Estivill X, Armengol L: **Accounting for uncertainty when assessing association between copy number and disease: a latent class model**. *BMC Bioinformatics* 2009, **10**:172.

14. Zheng J, Li Y, Abecasis G, Scheet P: **A comparison of approaches to account for uncertainty in analysis of imputed genotypes**. *Genet Epidemiol.* 2011, **35**(2):102–110.

15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet.* 2007, **81**(3):559–575, [http://pngu.mgh.harvard.edu/purcell/plink/].

16. Aulchenko Y, Struchalin M, van Duijn C: **ProbABEL package for genome-wide association analysis of imputed data**. *BMC Bioinformatics* 2010, **11**:134.

17. Subirana I, Diaz-Uriarte R, Lucas G, González J: **CNVassoc: Association analysis of CNV data using R**. *BMC Medical Genomics* 2011, **4**:47.

18. Subirana I, González J: **Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies**. *Genet Epidemiol.* 2013, **37**(5):465–477.

19. Cordell H: **Detecting gene-gene interactions that underlie human diseases**. *Nat Rev Genet.* 2009, **10**(6):392–404.

20. Lin HY D, Huang B: **Simple and efficient analysis of disease association with missing genotype data**. *Am J Hum Genet.* 2008, **82**(2):444–452.

21. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies via imputation of genotypes**. *Nat Genet.* 2007, **39**(7):906–913.

22. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS Genet.* 2009, **5**(6):e1000529.

23. Li WC Y and, Ding J, Scheet P, Abecasis G: **MACH: using sequence and genotype data to estimate haplotypes and unobserved genotypes**. *Genet Epidemiol.* 2010, **34**(8):816–834.

24. Lucas G, Lluís-Ganella C, Subirana I, Musameh M, Gonzalez J, Nelson C, Sentí M, O'Donnell C, Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, Schwartz S, Siscovick D, Melander O, Salomaa V, Purcell S, Altshuler D, Samani N, Kathiresan S, Elosua R: **Hypothesis-based analysis of gene-gene interactions and risk of myocardial infarction.** *PLoS One* 2012, **7**(8):e41730.

Figures

Figure 1 - Bias.

Estimated Hazard Ratio using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the cohort simulation study. Reference line (in grey) indicates no bias.

Figure 2 - Power.

Power (significance level of α=10e-6) using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the cohort simulation study.

Figure 3 - Coverage.

Observed coverage using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the cohort simulation study.

Figure 4 - QQ-plot.

Minus log10 p-values Quantile-Quantile plot corresponding to the 889,111 analyzed interactions using the three strategies (NAIVE, DOSAGE and LC) on the Framingham cohort data and taking time-to-coronary event as response variable.

Tables

Table 1 - Distribution of imputed SNPs according to combination of uncertainty (R^2) and minor allele frequency (MAF) from the Framingham cohort

	MAF (%)		
R^2 (%)	5 to 15	25 to 35	45 to 50
5 to 15	3,862	1,513	600
25 to 35	9,708	3,893	1,537
45 to 55	14,892	6,797	2,795
65 to 75	26,508	14,362	5,773
85 to 95	93,339	68,742	29,429

Additional Files

Additional file 1 — Likelihood, score and Hessian functions

This document provides formulas corresponding to likelihood functions for the LC model strategy as well as the first and second derivatives (Hessian matrix) in order to be able to perform the Newton-Raphson algorithm in obtaining parameter estimates and their standard errors.

Additional file 2 — Supplementary results

This document provides results for the case-control simulation study, as well as a table containing all simulated scenarios results from the cohort simulation study.

Figure 1

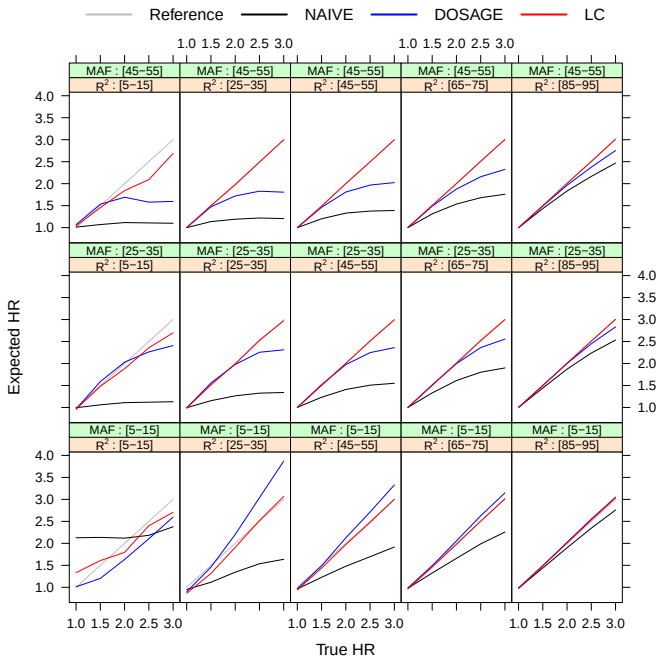


Figure 2

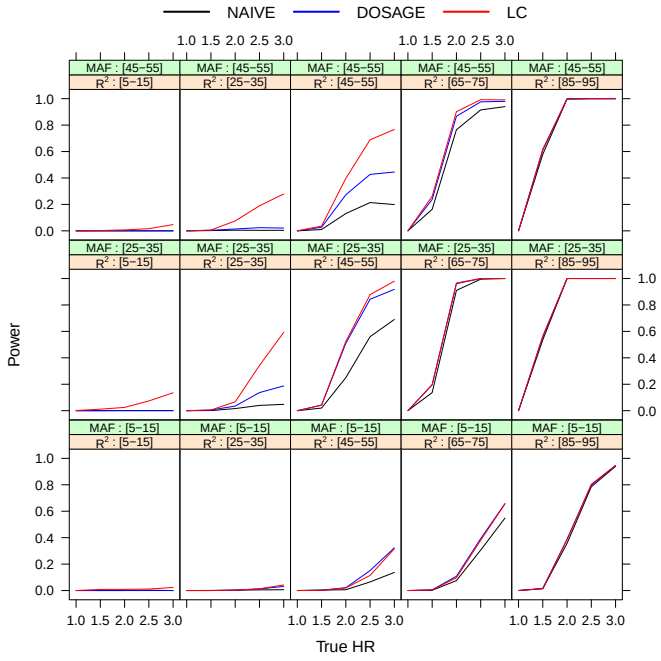


Figure 3

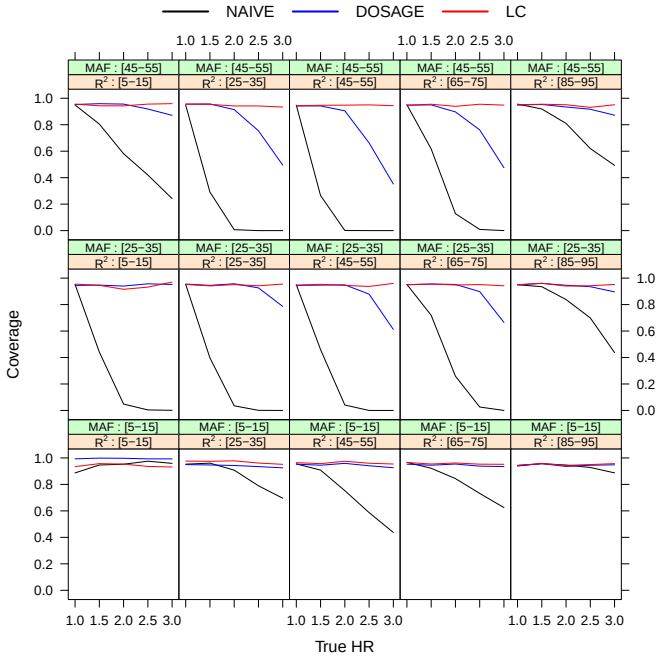
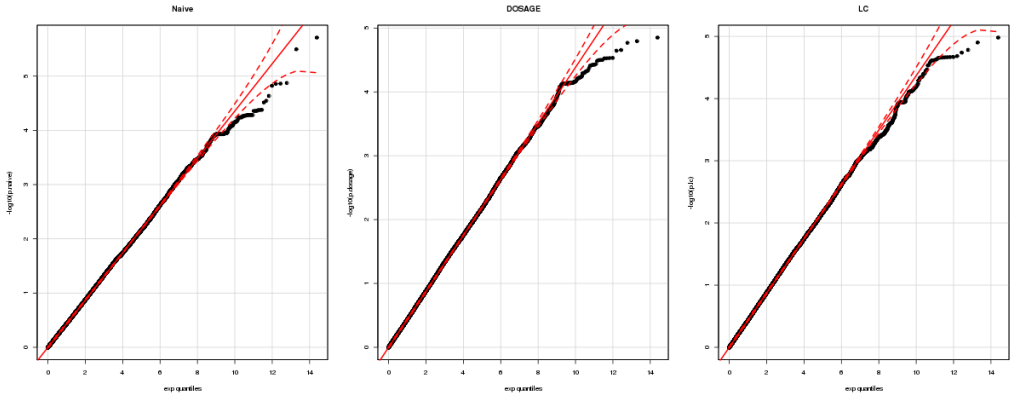


Figure 4



6.2. Material suplementario de las publicaciones

- **Artículo 1.** *Accounting for uncertainty when assessing association between copy number and disease: a latent class model.* (pág. 122).

- **Artículo 2.** *Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies.* (pág. 128)

- **Artículo 3.** *Interaction association analysis of imputed SNPs in case control and longitudinal studies.* (pág. 135)

- **Artículo 4.** *CNVassoc: Association analysis of CNV data using R.* (pág. 168). Se trata de la *vignette* del *package* `CNVassoc` disponible en el repositorio CRAN (http://www.r-project.org/).

Supplementary material of the paper: ***Accounting for uncertainty when assessing association between copy number and disease: a latent class model.*** Juan R. González, Isaac Subirana, Geòrgia Escaramís, Solymar Peraza, Alejandro Cáceres, Xavier Estivill, Lluís Armengol.

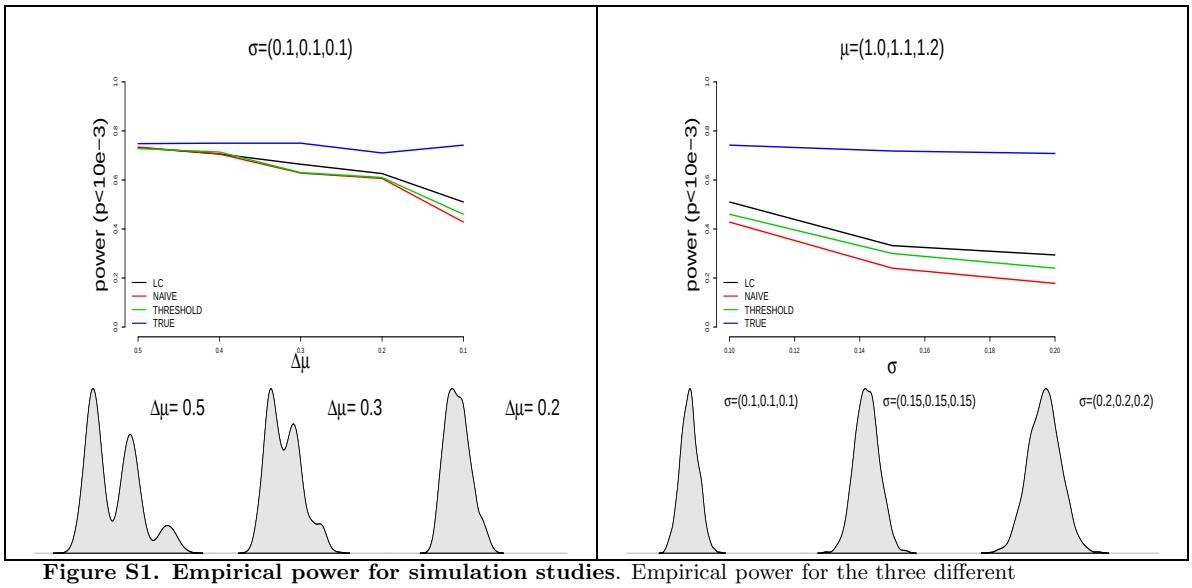


Figure S1. Empirical power for simulation studies. Empirical power for the three different approaches analyzed varying the quality of clustering for underlying copy number status. Left pannel is for a fixed set of variance and varying means, while the rigth pannel is for a fixed mean and varying variances.

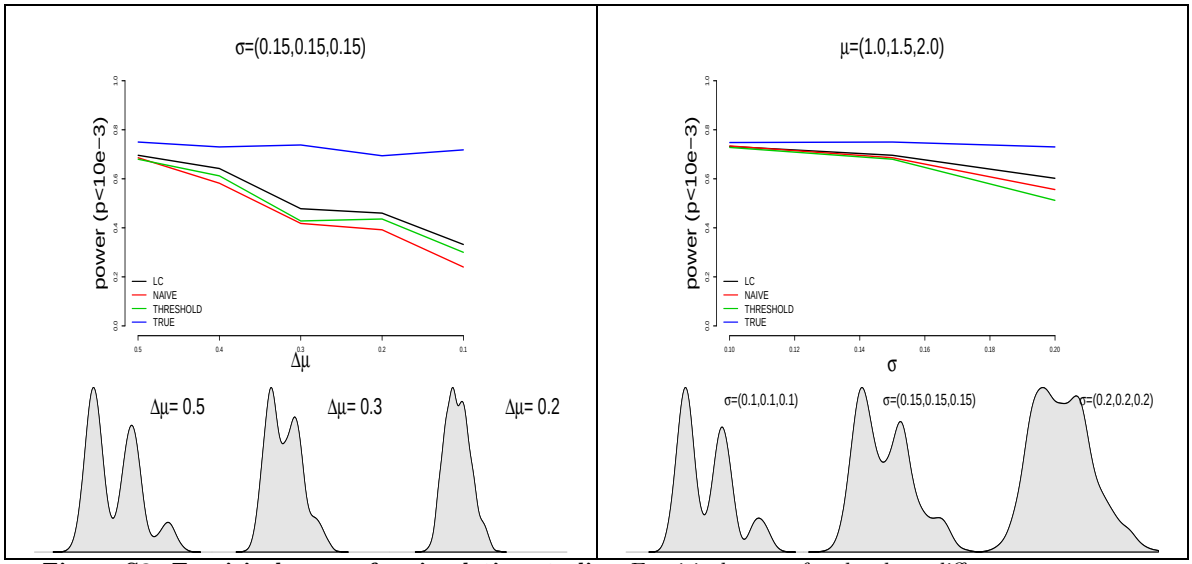


Figure S2. Empirical power for simulation studies. Empirical power for the three different approaches analyzed varying the quality of clustering for underlying copy number status. Left pannel is for a fixed set of variance and varying means, while the rigth pannel is for a fixed mean and varying variances.

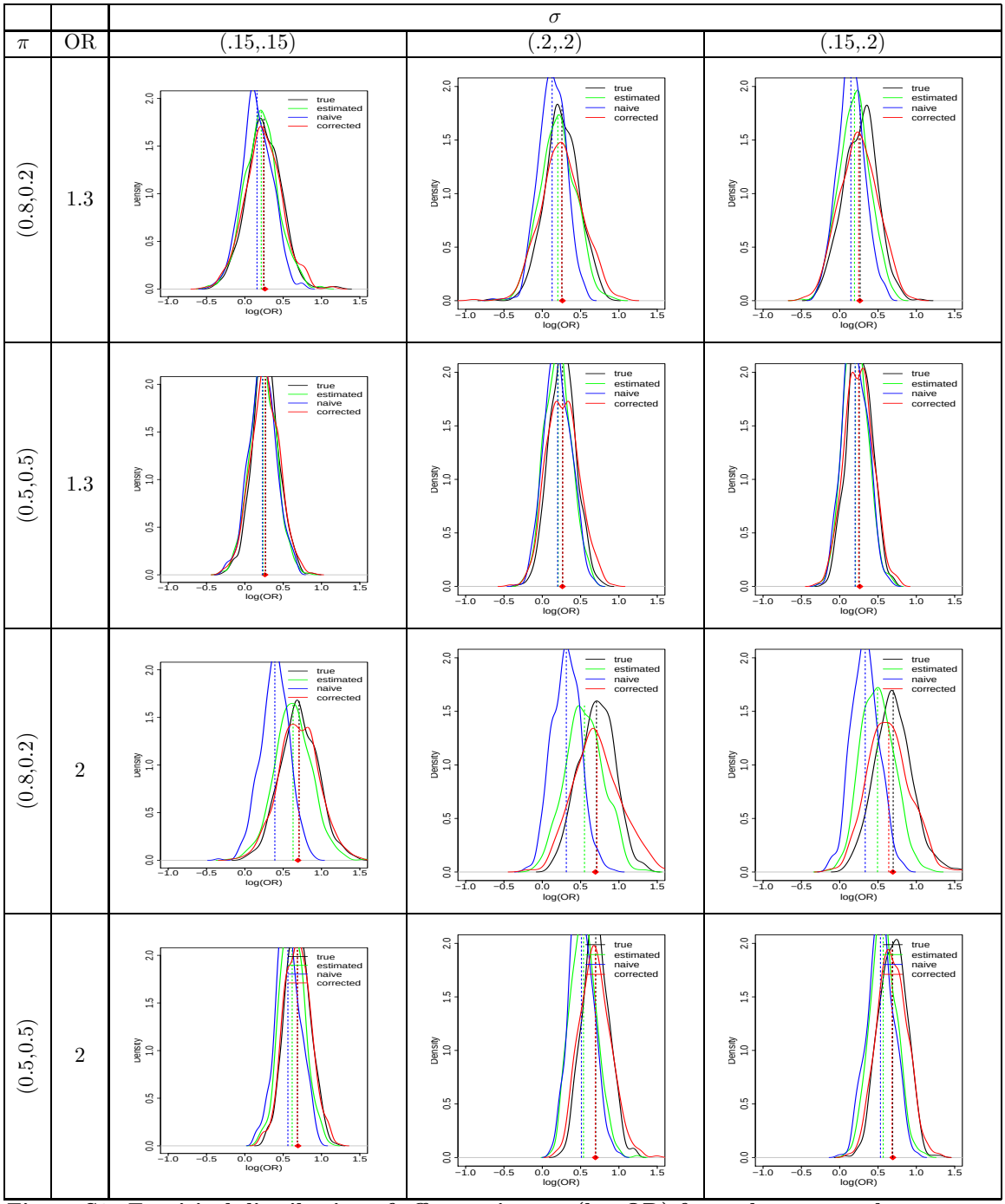


Figure S3. Empirical distribution of effect estimates (log OR) for each copy number status. Results for 1000 simulated case-control data sets (300/300), for different degrees of association (e.g. different OR) and different distributions of quantitative CNV measurements (e.g. varying clustering quality)

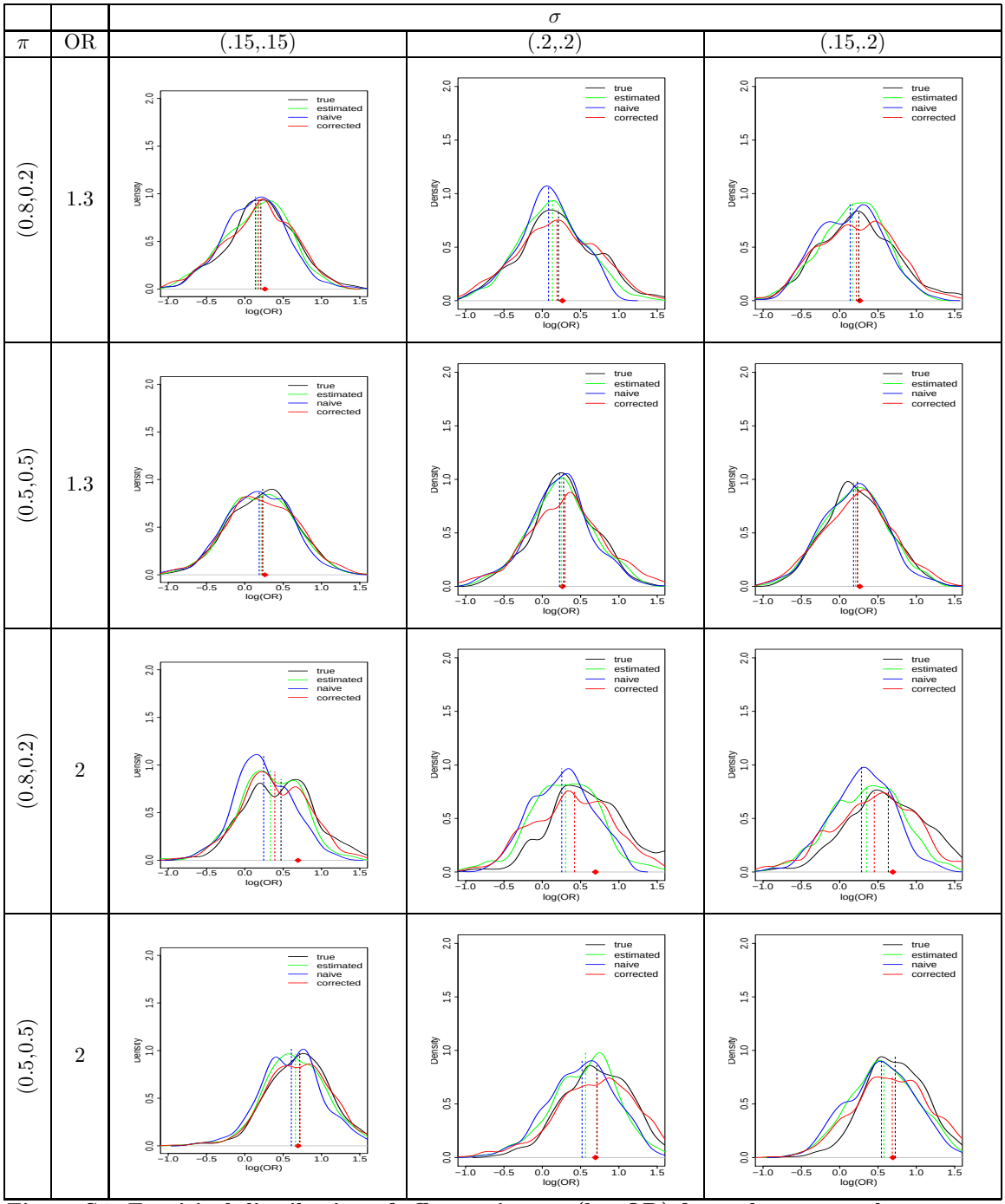


Figure S4. Empirical distribution of effect estimates (log OR) for each copy number status. Results for 1000 simulated case-control data sets (50/50), for different degrees of association (e.g. different OR) and different distributions of quantitative CNV measurements (e.g. varying clustering quality)

Table S1. Simulation study. Empirical coverage and power obtained in 1,000 simulations using the three different approaches, NAIVE, THRES and LC (see text for a description of each). Results are given for different scenarios, varying the number of individuals (I), the proportion of individuals with each copy number status (π), the odds ratio (e^{β}), and the variance for CNV quantitative measurements. The table also shows the variance of parameter estimates using the asymptotic (ASYM) variance compared with the empirical (EMP) variance.

				$\sigma_{\hat{\beta}}$		Coverage (%)				Power (%)			
I	π	e^{β}	σ	EMP	ASYM	SIM	NAIVE	THRES	LC	SIM	NAIVE	THRES	LC
50	0.8	1.3	(0.15,0.15)	0.5821	0.5898	94.2	96.2	95.8	96.8	6.6	5.4	6.4	4.6
50	0.8	1.3	(0.2,0.2)	0.5679	0.6605	93.0	94.0	93.0	96.2	5.2	4.8	4.2	3.6
50	0.8	1.3	(0.15,0.2)	0.5326	0.5846	96.6	96.2	95.4	97.4	6.8	4.8	4.8	3.0
50	0.8	2	(0.15,0.15)	0.6382	0.6512	94.2	92.6	89.0	94.0	22.0	16.8	11.2	15.4
50	0.8	2	(0.2,0.2)	0.6103	0.7057	92.8	92.2	82.8	95.2	16.8	9.4	7.4	7.0
50	0.8	2	(0.15,0.2)	0.6174	0.6407	95.6	87.0	79.4	93.0	19.4	10.6	9.8	9.6
50	0.5	1.3	(0.15,0.15)	0.4168	0.4367	94.0	94.2	95.2	93.8	11.6	10.0	9.2	10.0
50	0.5	1.3	(0.2,0.2)	0.4298	0.4838	94.6	93.8	94.0	95.4	12.6	7.0	7.0	7.2
50	0.5	1.3	(0.15,0.2)	0.3984	0.4578	95.2	95.2	95.2	95.6	11.4	8.6	9.4	8.2
50	0.5	2	(0.15,0.15)	0.4231	0.4495	95.6	94.6	93.8	94.6	39.4	32.4	32.4	32.6
50	0.5	2	(0.2,0.2)	0.4022	0.5020	97.0	95.0	94.6	98.2	42.2	23.8	23.2	25.2
50	0.5	2	(0.15,0.2)	0.4345	0.4696	94.4	93.4	94.4	95.6	47.4	30.8	29.8	33.4
300	0.8	1.3	(0.15,0.15)	0.2291	0.2341	94.0	94.0	89.2	93.2	20.4	15.2	17.0	17.8
300	0.8	1.3	(0.2,0.2)	0.2208	0.2667	94.6	94.4	88.6	96.4	23.0	17.0	11.0	16.2
300	0.8	1.3	(0.15,0.2)	0.2192	0.2373	94.2	93.6	89.2	96.0	23.4	15.8	13.2	18.0
300	0.8	2	(0.15,0.15)	0.2452	0.2610	94.2	93.6	66.0	94.6	85.4	78.4	58.6	79.2
300	0.8	2	(0.2,0.2)	0.2334	0.2996	95.8	89.8	43.2	96.0	84.2	60.8	42.6	66.6
300	0.8	2	(0.15,0.2)	0.2455	0.2591	93.8	83.0	43.8	94.6	85.8	62.8	44.8	67.4
300	0.5	1.3	(0.15,0.15)	0.1711	0.1775	93.6	93.8	94.0	93.8	37.0	30.8	31.2	32.4
300	0.5	1.3	(0.2,0.2)	0.1709	0.1970	94.4	93.8	92.8	93.6	36.6	24.4	25.0	28.2
300	0.5	1.3	(0.15,0.2)	0.1582	0.1866	96.8	95.2	94.4	95.2	34.6	22.8	24.6	25.2
300	0.5	2	(0.15,0.15)	0.1621	0.1823	95.8	95.2	90.4	95.8	98.4	96.8	93.0	97.2
300	0.5	2	(0.2,0.2)	0.1692	0.2030	96.2	84.0	82.4	96.0	99.2	89.4	90.0	94.2
300	0.5	2	(0.15,0.2)	0.1793	0.1904	95.4	88.2	83.0	95.2	98.2	92.4	88.2	94.4

Table S2. Simulation study. Empirical coverage and power obtained in 1,000 simulations using the three different approaches: NAIVE, THRES and LC (read text to have a description of each one. LCa means LC using Newton-Raphson procedure and LCb is LC using bootstrap approach). The results are given for different scenarios varying number of individuals (I), proportion of individuals in each copy number status (pi), odds ratio (e^{β}) and variance for CNV quantitative measurements. The table also shows the variance of parameter estimates using the asymptotic (ASYM) variance and variance obtained using bootstrap procedure (BOOT) compared with the empirical (EMP) variance.

				σ_{β}			Coverage (%)					Power (%)				
n	π	e^{β}	σ	EMP	ASYM	BOOT	SIM	NAIVE	THRES	LCa	LCb	SIM	NAIVE	THRES	LCa	LCb
50	0.8	1.3	(0.15,0.15)	0.4539	0.5629	0.6234	66.4	66.4	65.4	66.8	66.2	1.6	1.2	1.6	1.2	1.4
50	0.8	1.3	(0.2,0.2)	0.5064	0.5985	0.6730	58.0	58.0	57.0	59.4	58.2	2.6	1.4	0.6	0.8	1.2
50	0.8	1.3	(0.15,0.2)	0.5412	0.5547	0.5991	76.2	75.8	74.6	76.6	76.2	4.2	1.8	3.6	1.6	2.4
50	0.8	2	(0.15,0.15)	0.4677	0.5886	0.6609	43.0	42.0	39.2	42.8	42.8	3.6	1.2	1.2	1.4	1.8
50	0.8	2	(0.2,0.2)	0.5105	0.6312	0.7813	46.4	43.6	39.0	45.8	45.8	6.4	2.2	3.0	2.6	4.6
50	0.8	2	(0.15,0.2)	0.5589	0.6021	0.7131	66.8	61.8	58.4	65.6	65.2	10.4	5.6	4.0	4.6	6.4
50	0.5	1.3	(0.15,0.15)	0.4357	0.4377	0.4517	94.0	93.8	94.0	94.0	93.2	10.2	9.4	7.6	7.6	8.4
50	0.5	1.3	(0.2,0.2)	0.4042	0.4864	0.5055	93.8	92.6	93.2	93.0	92.0	12.0	9.6	9.0	8.8	10.4
50	0.5	1.3	(0.15,0.2)	0.4180	0.4572	0.4765	95.0	95.2	94.6	95.6	94.4	10.2	7.6	5.6	7.4	8.8
50	0.5	2	(0.15,0.15)	0.4134	0.4500	0.4682	95.0	95.2	95.8	94.8	93.0	42.4	36.4	34.0	34.6	37.2
50	0.5	2	(0.2,0.2)	0.4461	0.5010	0.5272	91.8	89.8	90.4	91.6	90.8	42.0	29.0	27.4	30.0	33.4
50	0.5	2	(0.15,0.2)	0.4059	0.4670	0.4860	94.2	92.0	90.8	94.0	91.4	42.0	31.0	30.6	32.4	35.2
300	0.8	1.3	(0.15,0.15)	0.2167	0.2357	0.2381	95.6	96.4	91.6	96.4	95.6	23.8	18.2	14.6	19.8	23.0
300	0.8	1.3	(0.2,0.2)	0.2001	0.2680	0.2717	96.6	95.0	89.0	96.2	94.6	23.2	16.2	10.8	15.8	19.4
300	0.8	1.3	(0.15,0.2)	0.2132	0.2371	0.2400	95.8	94.4	90.2	95.0	94.0	23.2	14.6	12.8	16.6	19.2
300	0.8	2	(0.15,0.15)	0.2398	0.2592	0.2644	95.2	94.2	64.0	96.0	94.2	85.8	74.6	55.2	78.2	79.6
300	0.8	2	(0.2,0.2)	0.2469	0.2963	0.3065	93.6	87.2	38.2	95.6	94.0	86.0	56.0	39.2	63.2	66.0
300	0.8	2	(0.15,0.2)	0.2395	0.2589	0.2633	94.4	82.8	42.2	94.4	92.8	86.2	61.0	42.8	65.0	68.6
300	0.5	1.3	(0.15,0.15)	0.1580	0.1774	0.1779	95.8	96.2	94.4	95.8	94.6	35.6	28.6	27.6	29.8	32.4
300	0.5	1.3	(0.2,0.2)	0.1742	0.1967	0.1968	93.0	92.0	92.0	92.8	91.6	38.6	27.2	23.4	28.2	31.6
300	0.5	1.3	(0.15,0.2)	0.1686	0.1864	0.1878	94.0	93.8	92.4	94.6	94.2	36.6	25.6	25.2	26.6	29.2
300	0.5	2	(0.15,0.15)	0.1642	0.1825	0.1834	96.4	94.2	89.4	95.2	94.2	99.0	97.0	93.8	97.4	97.8
300	0.5	2	(0.2,0.2)	0.1681	0.2033	0.2054	95.4	86.8	80.4	94.4	92.4	98.6	90.4	88.2	94.6	94.4
300	0.5	2	(0.15,0.2)	0.1647	0.1903	0.1911	96.2	89.4	84.4	94.2	92.4	98.8	93.6	91.2	95.4	95.4

Additional file: Supplementary figures

Genetic association analysis and meta-analysis of imputed SNPs in longitudinal studies

Isaac Subirana[1,2,3], Juan R González[4,5,1*]

1 CIBER Epidemiology and Public Health (CIBERESP), Spain

2 Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, IMIM, Parc de Salut Mar, Spain

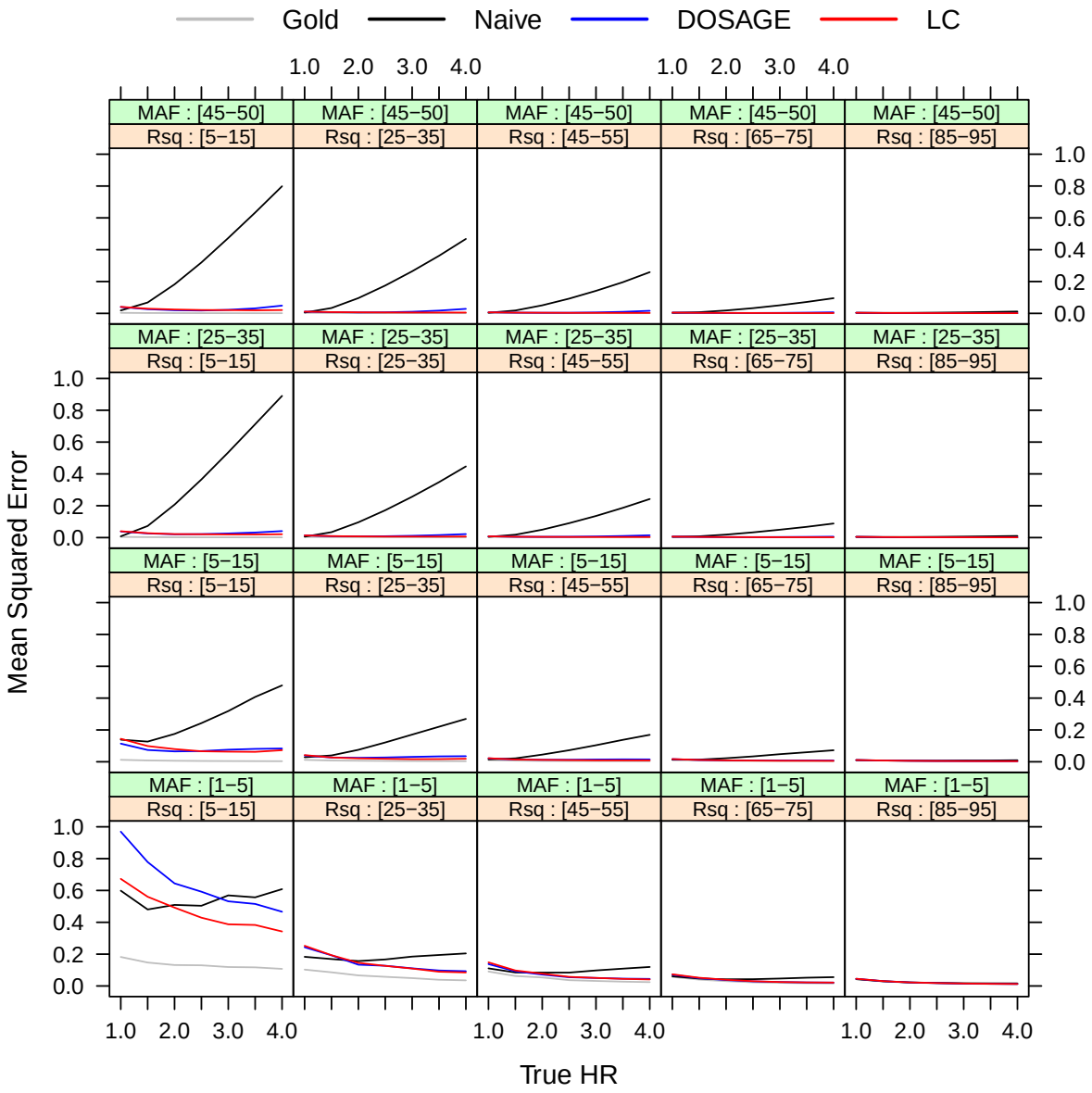
3 Statistics Department, University of Barcelona, Spain

4 Center for Research in Environmental Epidemiology (CREAL), Spain

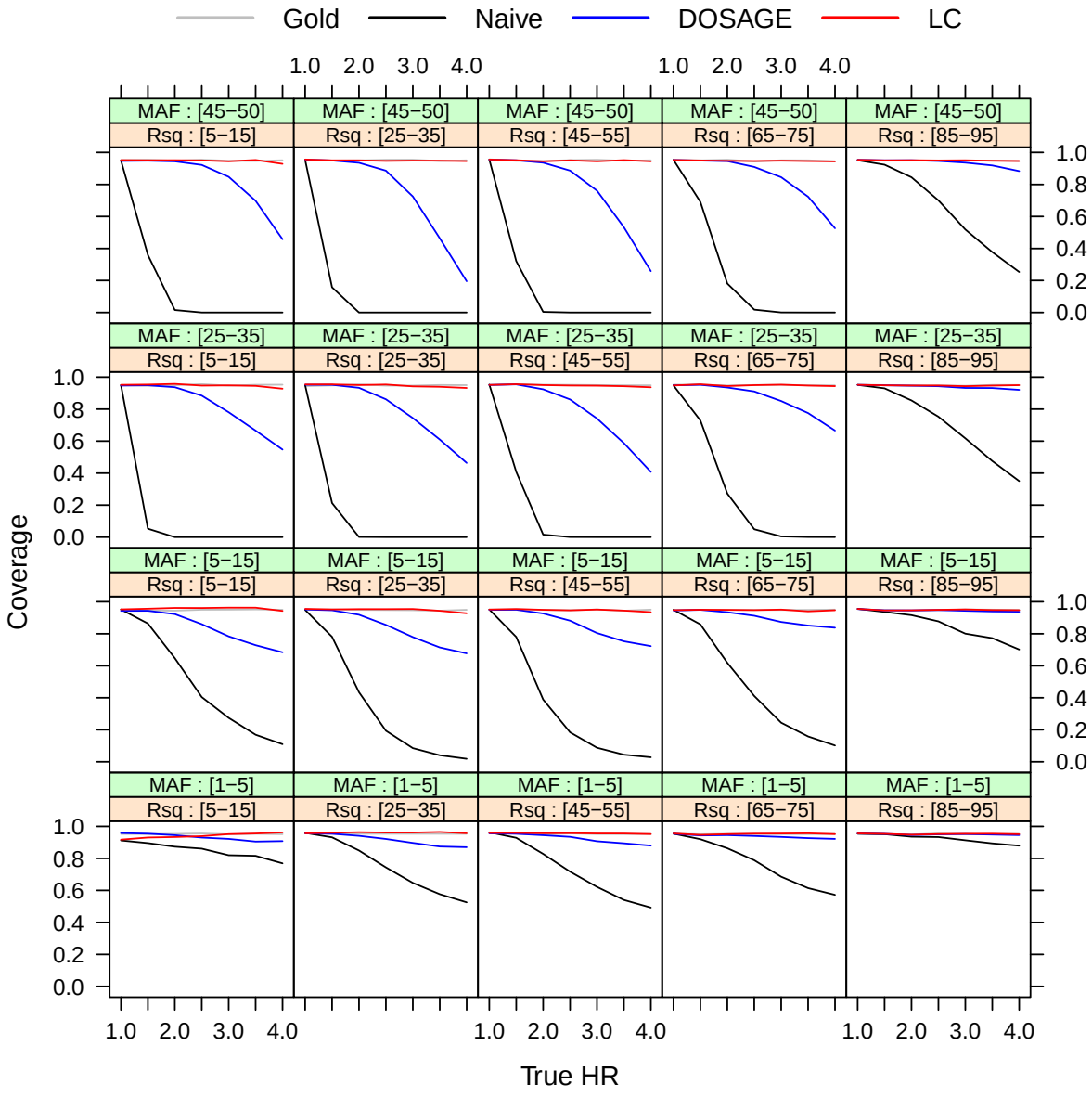
5 Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Spain

* Corresponding author: Juan R Gonzalez (jrgonzalez@creal.cat)

Simulation study with censored Weibull response

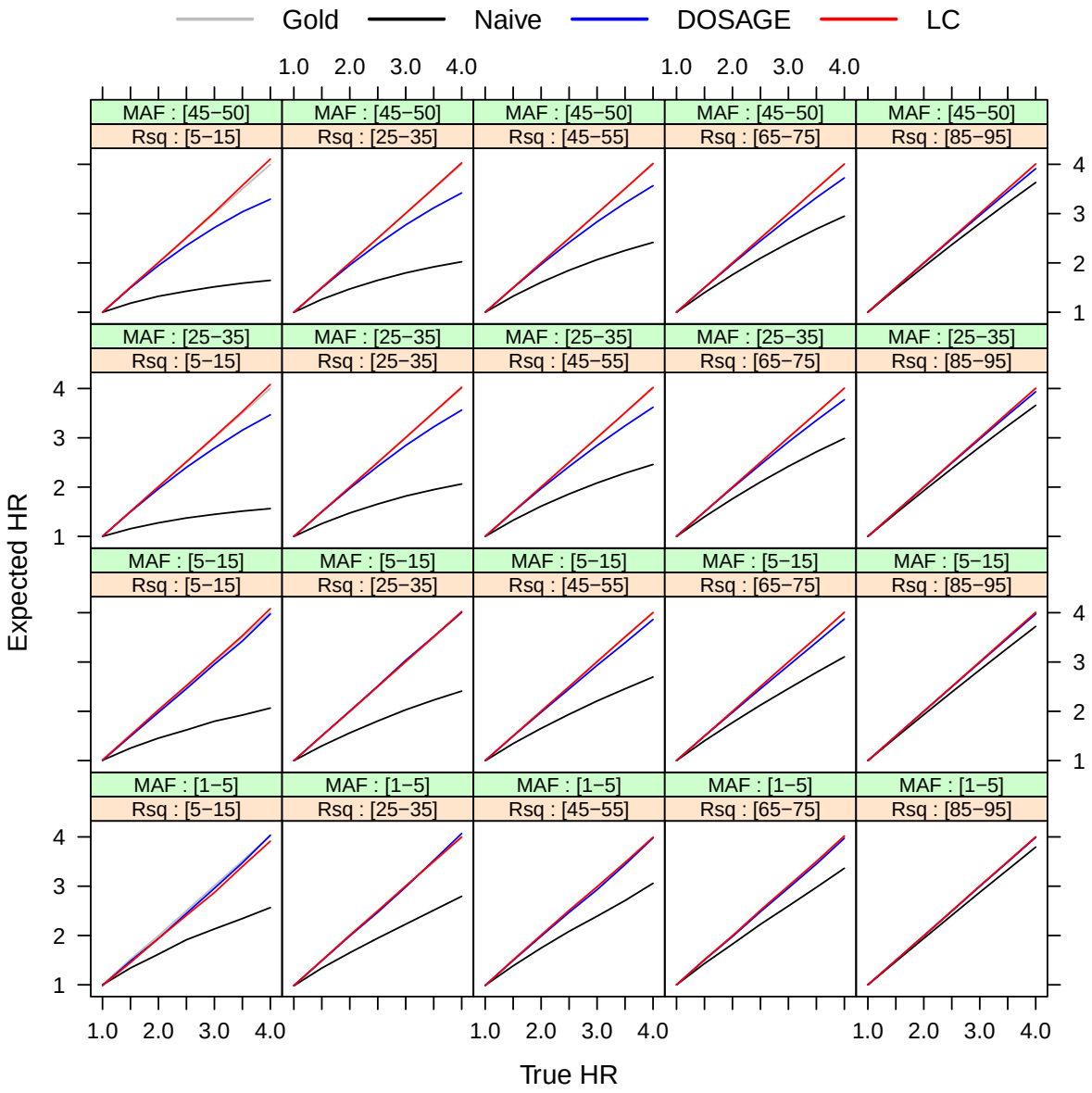


Supplementary Figure 1: Accuracy measured as Mean Squared Error (MSE) according to minor allele frequency and uncertainty (R^2).

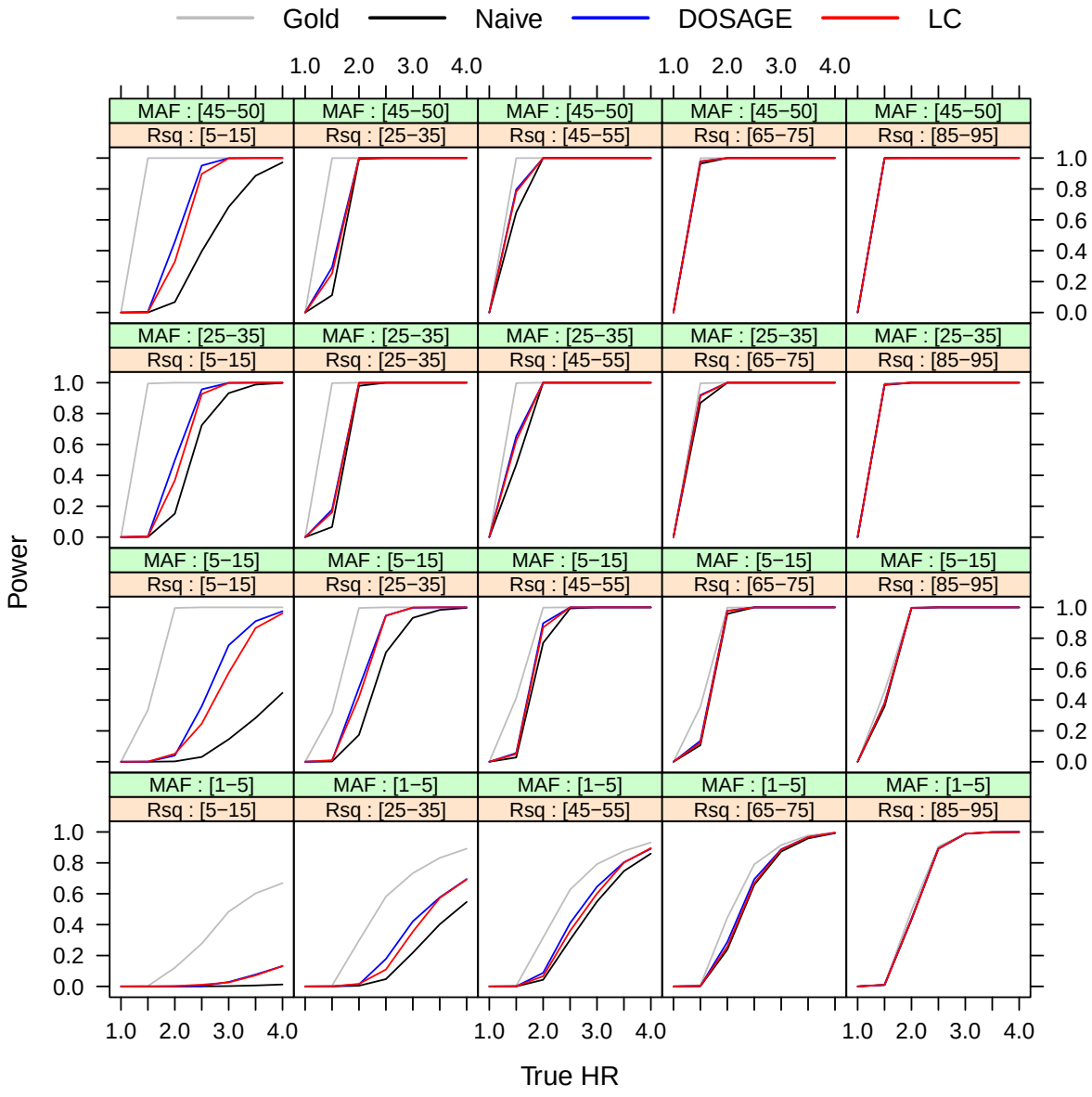


Supplementary Figure 2: 95% confidence interval coverage according to minor allele frequency and uncertainty (R^2).

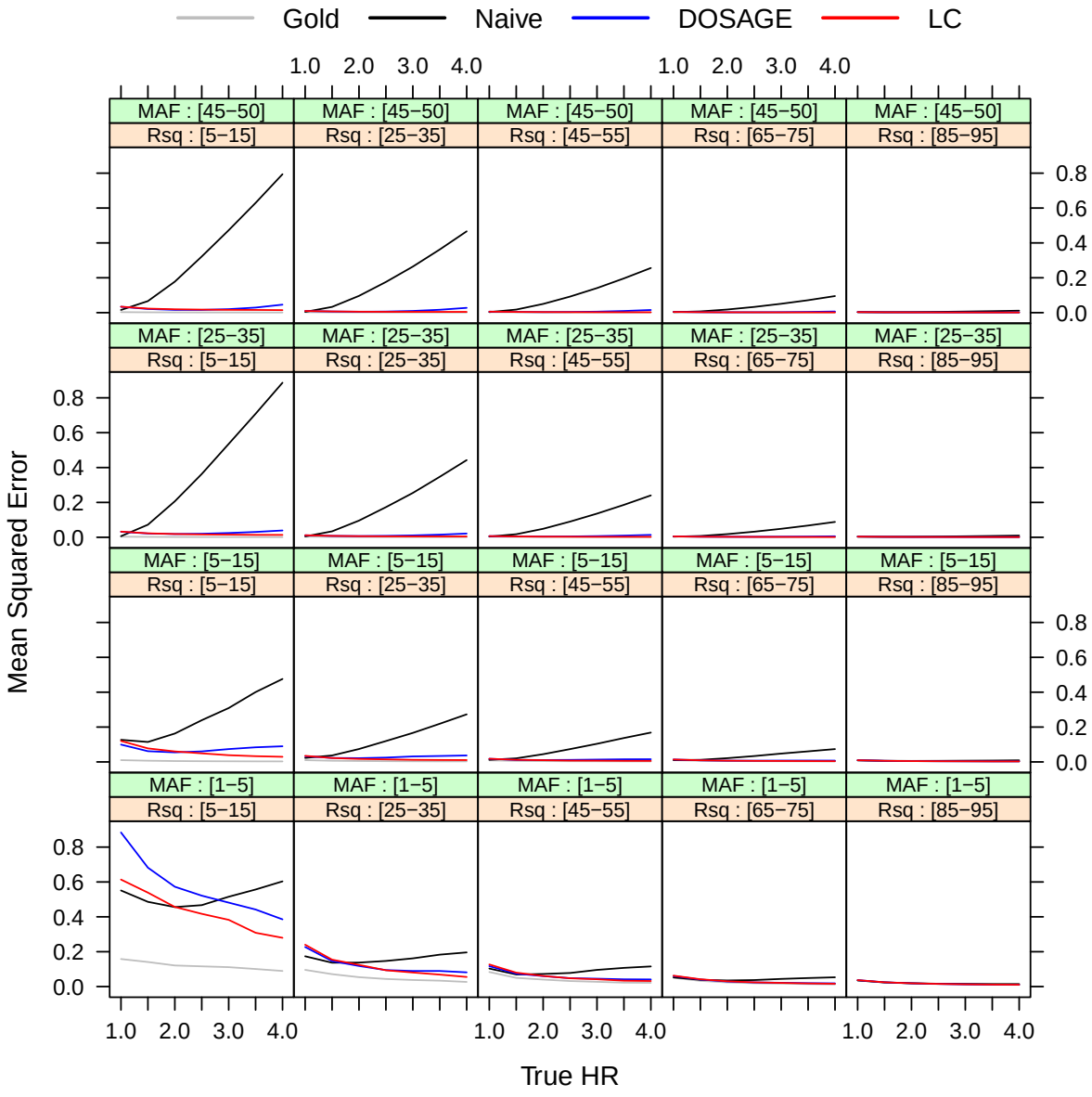
Simulation study with empirical response



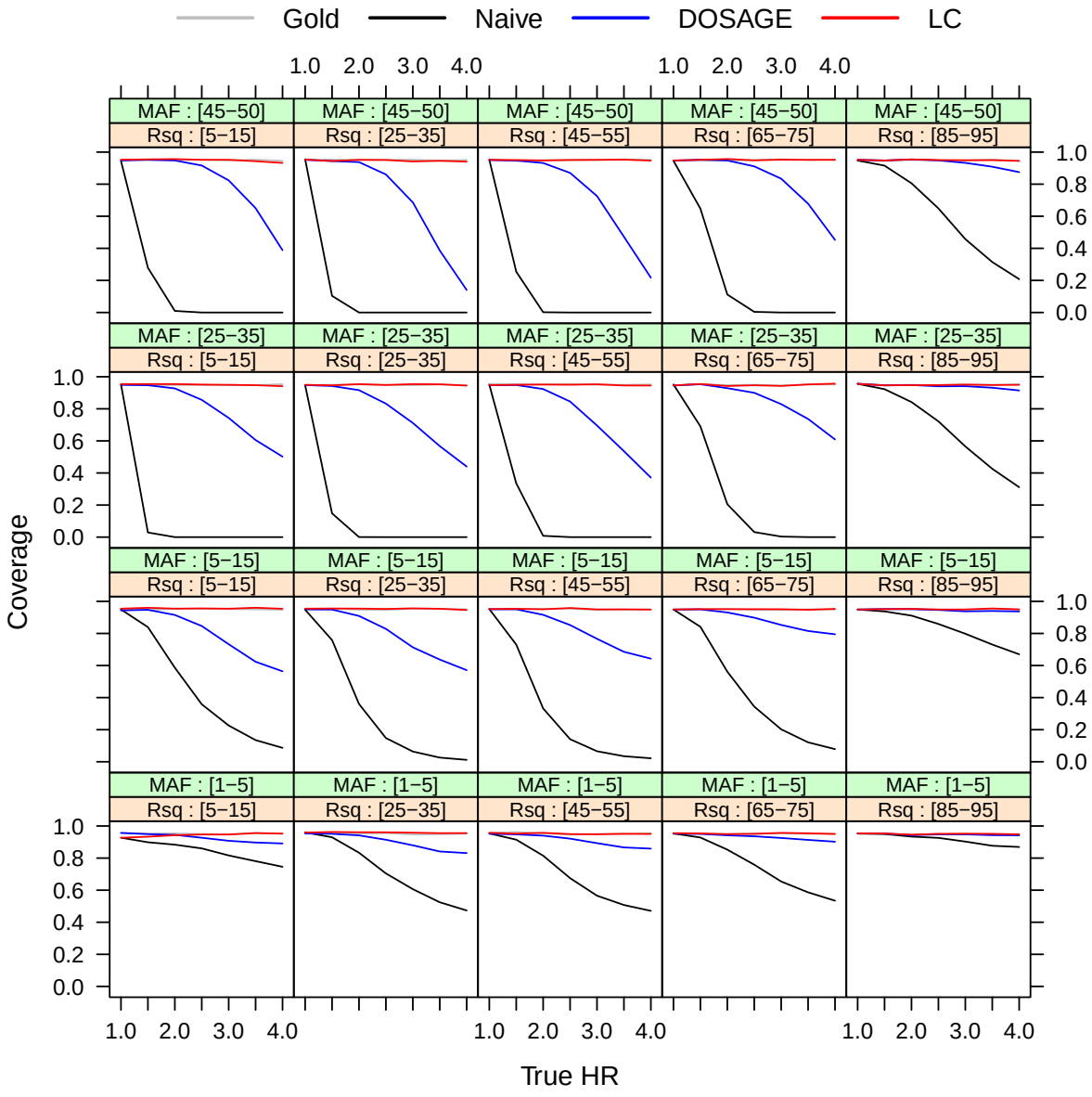
Supplementary Figure 3: Hazard ratio according to minor allele frequency (MAF) and uncertainty (R^2).



Supplementary Figure 4: Power according to minor allele frequency (MAF) and uncertainty (R^2), taking GWA significance level, α = 2·10e-8.



Supplementary Figure 5: Accuracy measured as Mean Squared Error (MSE) according to minor allele frequency and uncertainty (R^2).



Supplementary Figure 6: 95% confidence interval coverage according to minor allele frequency and uncertainty (R^2).

Additional file 1: Likelihood, score and Hessian functions

Interaction association analysis of imputed SNPs in case control and longitudinal studies

Isaac Subirana[1,2,3], Juan R González[4,1,5*]

1 CIBER Epidemiology and Public Health (CIBERESP), Spain

2 IMIM, Parc de Salut Mar, Spain

3 Statistics Department, University of Barcelona, Spain

4 Center for Research in Environmental Epidemiology (CREAL), Spain

5 Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Spain

* Corresponding author: Juan R Gonzalez (jrgonzalez@creal.cat)

The logarithm of the likelihood function (log-likelihood) for the Latent Class model is

$$logL(\boldsymbol{Y};\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=0}^{2} \sum_{l=0}^{2} \text{Prob}(Y_i|\, \text{SNP}_1 = k, \text{SNP}_1 = l, \boldsymbol{C}_i; \boldsymbol{\Theta}) w_{ikl} \right\}$$

where $\boldsymbol{\Theta}$ is the parameter vector. Finally, w_{ikl} is the joint probability of having k and l risk alleles for the first and second SNP, respectively:

$$w_{ikl} \equiv \text{Prob}(\text{SNP}_1 = k, \text{SNP}_2 = l|i) = \text{Prob}(\text{SNP}_1 = k|i)\, \text{Prob}(\text{SNP}_2 = l|i)$$

For convenience, we will denote θ_s as the s-th component of $\boldsymbol{\Theta}$ vector. The first derivatives of the Log-likelihood function with respect to the s-th component of $\boldsymbol{\Theta}$:

$$\frac{\partial \log L(\boldsymbol{Y};\boldsymbol{\Theta})}{\partial \theta_s} = \sum_{i=1}^{n} \frac{\sum_{k=0}^{2} \sum_{l=0}^{2} \frac{\partial h_{ikl}}{\partial \theta_s}}{g_i}$$

and the second derivatives of the log-likelihood function:

$$\frac{\partial^2 \log L(\boldsymbol{Y};\boldsymbol{\Theta})}{\partial \theta_s \partial \theta_{s'}} = \sum_{i=1}^{n} \frac{\sum_{k=0}^{2} \sum_{l=0}^{2} \frac{\partial h_{ikl}}{\partial \theta_s \partial \theta_{s'}} g_i - \sum_{k=0}^{2} \sum_{l=0}^{2} \frac{\partial h_{ikl}}{\partial \theta_s} \sum_{k=0}^{2} \sum_{l=0}^{2} \frac{\partial h_{ikl}}{\partial \theta_{s'}}}{g_i^2}$$

where

$$g_i \equiv \sum_{k=0}^{2}\sum_{l=0}^{2} h_{ikl}$$

and

$$h_{ikl} \equiv \text{Prob}(Y_i|\,\text{SNP}_1 = k, \text{SNP}_2 = l, \boldsymbol{C}_i; \boldsymbol{\Theta}) w_{ikl},$$

Finally, we define the linear predictor as

$$\eta_{ikl} \equiv \beta_0 + \beta_1 k + \beta_2 l + \beta_{12} kl + \boldsymbol{C}_i \boldsymbol{\gamma}\prime$$

where, β_0 is the constant, β_1 and β_2 are the main effects, β_{12} is the interaction effect, $\boldsymbol{C}_i$ is the covariate vector, $\boldsymbol{\gamma}$ is the covariate coefficients parameter vector and k and l are the number of risk alleles (0, 1 or 2) for the first and second SNP, respectively.

Case control study

For case-control study, the h_{ikl} function takes the form

$$h_{ikl} = w_{ikl} \frac{e^{y_i \eta_{ikl}}}{1 + e^{\eta_{ikl}}}$$

The vector of parameters, $\boldsymbol{\Theta}$, consists of the constant, β_0, the main effects log-Odds Ratio for SNP_1 and SNP_2, β_1 and β_2, the log-Odds Ratio for the interaction effect, β_{12}, and the log-Odds Ratio for the covariates, γ_j.

The first and second derivatives of the function h_{ikl} are:

- First derivatives:

$$\frac{\partial h_{ikl}}{\partial \theta_s} = x_{ik} h_{ikl} (y_i - p_{ikl})$$

- Second derivatives:

$$\frac{\partial^2 h_{ikl}}{\partial \theta_s \partial \theta_{s'}} = \frac{\partial h_{ikl}}{\partial \theta_s} x_{is} (1 - p_{ikl}) - h_{ikl} x_{is'} p_{ikl} (1 - p_{ikl})$$

The term p_{ikl} is the probability of being a case, $Y_i = 1$, given that the number of risk alleles is k and l for the first and second SNP, respectively:

$$p_{ikl} \equiv \text{Prob}(Y_i = 1 | \, \text{SNP}_1 = k, \text{SNP}_2 = l; \boldsymbol{\Theta}) = \frac{1}{1 + e^{-\eta_{ikl}}}$$

Cohort study

For cohort studies, two situations must be considered: whether the time to response for the i-th individual has been observed ($\delta_i = 1$) or not ($\delta_i = 0$), where δ is called the censor indicator or variable.

Censored observations:

h_{ikl} function takes the form:

$$h_{ikl} \equiv w_{ikl} e^{-\lambda_{ikl} y_i^{\phi}}$$

where $\lambda_{ikl} = e^{\eta_{ikl}}$.

The first and second derivatives of the function h_{ikl} are:

- First derivatives
  - with respect to any parameter but ϕ:

$$\frac{\partial h_{ikl}}{\partial \theta_s} = h_{ikl} x_{is} (-y_i^{\phi} \lambda_{ikl})$$

  - with respect to the shape parameter ϕ:

$$\frac{\partial h_{ikl}}{\partial \phi} = -h_{ikl} \lambda_{ikl} y_i^{\phi} \log(y_i)$$

- Second derivatives:
  - with respect to any parameter but ϕ:

$$\frac{\partial^2 h_{ikl}}{\partial \theta_s \partial \theta_{s'}} = x_{is} \left[-y_i^{\phi} \lambda_{ikl} \left(x_{is'} h_{ikl} - \frac{\partial h_{ikl}}{\partial \theta_{s'}} \right) \right]$$

- with respect to shape parameter ϕ:

$$\frac{\partial^2 h_{ikl}}{\partial \phi^2} = -\lambda_{ikl} \log(y_i) y_i^{\phi} \left(\frac{\partial h_{ikl}}{\partial \phi} + h_{ikl} \log(y_i) \right)$$

- Cross-derivatives between the constant/coefficients ($\theta_s \in \{\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_j\}$) and the shape parameter ϕ

$$\frac{\partial^2 h_{ikl}}{\partial \theta_s \partial \phi} = x_{is} \left[-\lambda_{ikl} y_i^{\phi} \left(\frac{\partial h_{ikl}}{\partial \phi} + h_{ikl} \log(y_i) \right) \right]$$

Non censored observations:

h_{ikl} function takes the form:

$$h_{ikl} = w_{ikl} \lambda_{ikl} \phi y_i^{\phi - 1} e^{-\lambda_{ikl} y_i^{\phi}}$$

The first and second derivatives of the function h_{ikl} are:

- First derivatives

  - with respect to any parameter but ϕ:

$$\frac{\partial h_{ikl}}{\partial \theta_s} = h_{ikl} x_{is} (1 - y_i^{\phi} \lambda_{ikl})$$

  - with respect to the shape parameter ϕ:

$$\frac{\partial h_{ikl}}{\partial \phi} = -h_{ikl} \left(\lambda_{ikl} y_i^{\phi} \log(y_i) - \log(y_i) - \frac{1}{\phi} \right)$$

- Second derivatives:

  - with respect to any parameter but ϕ:

$$\frac{\partial^2 h_{ikl}}{\partial \theta_s \partial \theta_{s'}} = x_{is} \left[\frac{\partial h_{ikl}}{\partial \theta_{s'}} \left(1 - \lambda_{ikl} y_i^{\phi} \right) - h_{ikl} \lambda_{ikl} y_i^{\phi} x_{is'} \right]$$

– with respect to shape parameter ϕ:

$$\frac{\partial^2 h_{ikl}}{\partial \phi^2} = -\frac{\partial h_{ikl}}{\partial \phi}\left(y_i^{\phi}\log(y_i)\lambda_{ikl} - \log(y_i) - \frac{1}{\phi}\right) - h_{ikl}\left(y_i^{\phi}\log^2(y_i)\lambda_{ikl} + \frac{1}{\phi^2}\right)$$

- Cross-derivatives between the constant/coefficients ($\theta_s \in \{\beta_0, \beta_1, \beta_2, \beta_{12}, \gamma_j\}$) and the shape parameter ϕ

$$\frac{\partial^2 h_{ikl}}{\partial \theta_s \partial \phi} = x_{is}\left[\frac{\partial h_{ikl}}{\partial \phi}\left(1 - \lambda_{ikl}y_i^{\phi}\right) - h_{ikl}\lambda_{ikl}y_i^{\phi}\log(y_i)\right]$$

For all previous expressions, including case-control and cohort studies, the term x_{is} takes the value of 1, k, l, kl or C_{ij} depending on whether θ_s is equal to β_0, β_1, β_{12} or γ_j, respectively. Similarly, $x_{is'}$ takes the value of 1, k, l, kl or C_{ij} depending on whether $\theta_{s'}$ is equal to β_0, β_1, β_{12} or γ_j respectively. And the term C_{ij} is the value of j-th covariate for the i-th individual.

Additional file 2: Supplementary results

Interaction association analysis of imputed SNPs in case control and longitudinal studies

Isaac Subirana[1,2,3], Juan R González[4,1,5*]

1 CIBER Epidemiology and Public Health (CIBERESP), Spain

2 IMIM, Parc de Salut Mar, Spain

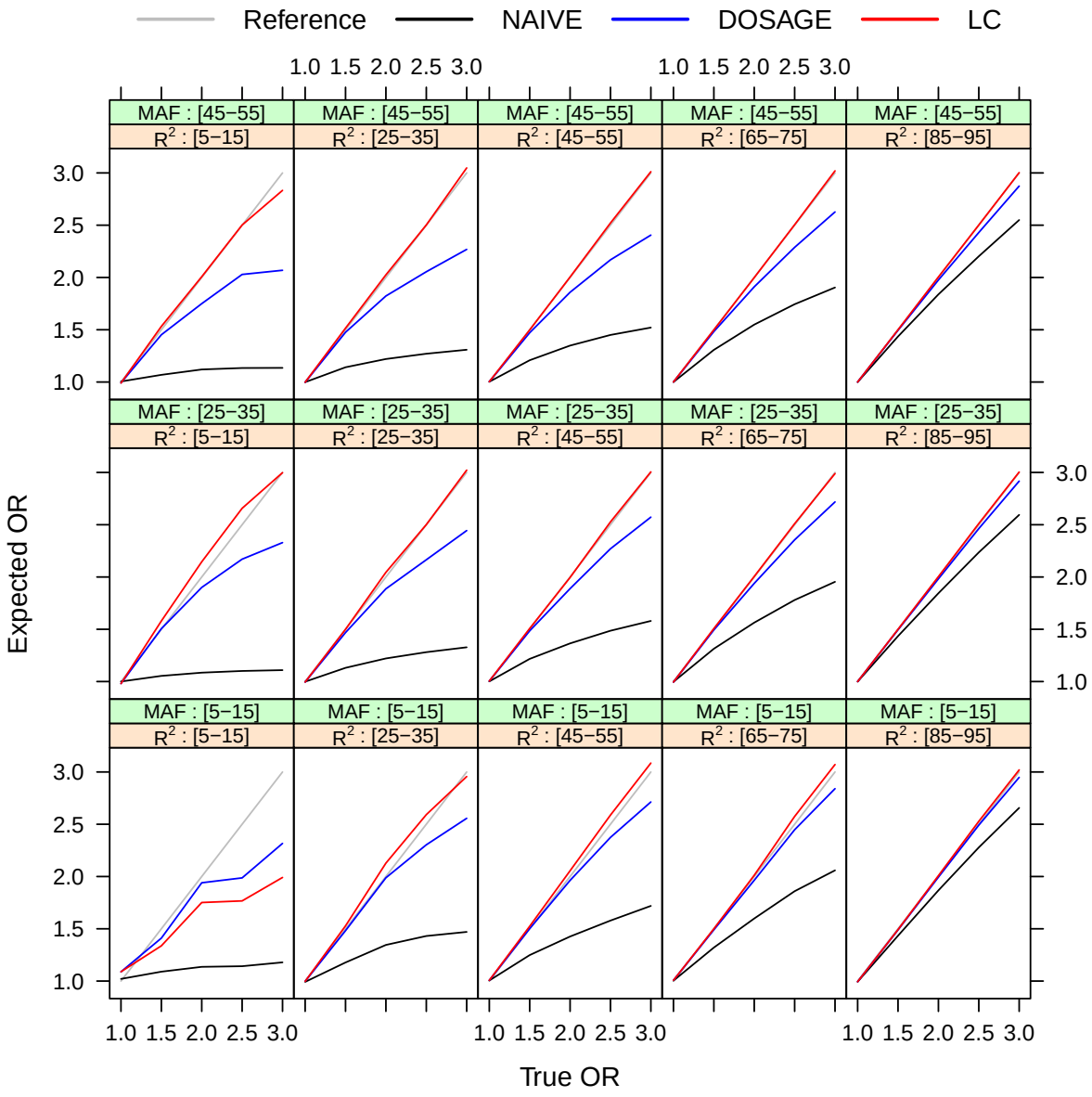
3 Department of Statistics, University of Barcelona, Spain

4 Center for Research in Environmental Epidemiology (CREAL), Spain

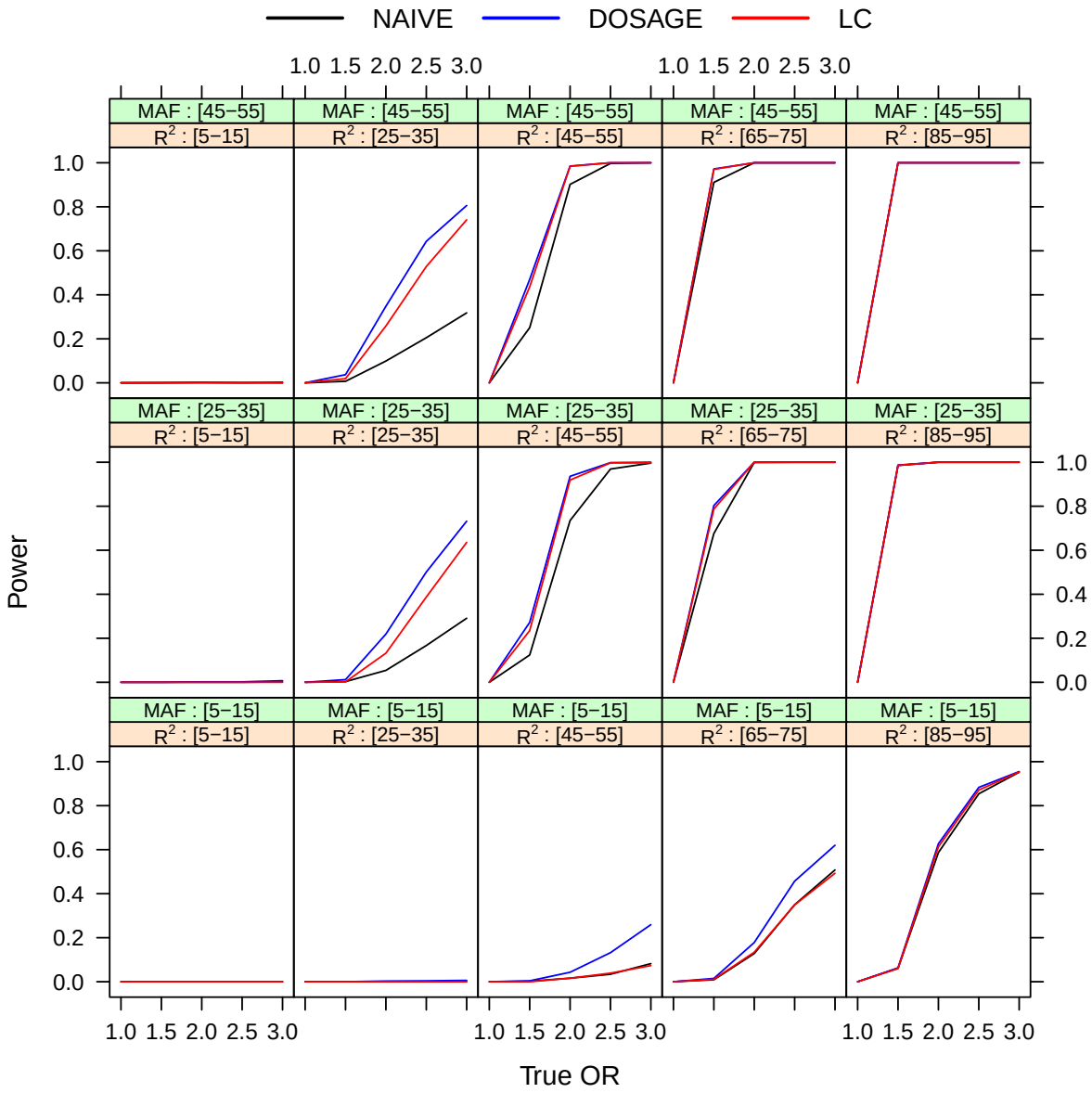
5 Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Spain

* Corresponding author: Juan R Gonzalez (jrgonzalez@creal.cat)

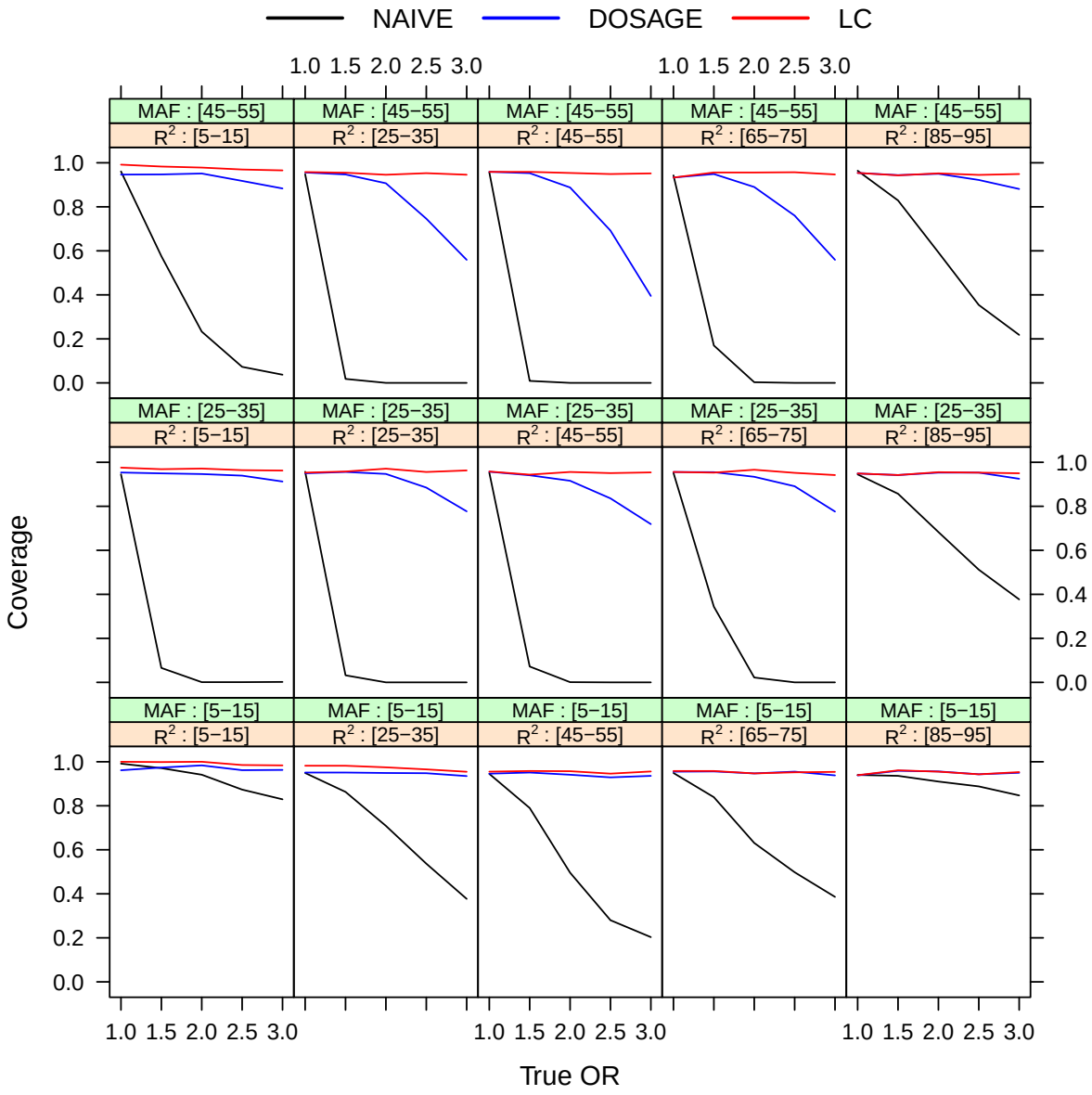
Supplementary Figure 1: Estimated Odds Ratio using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the case control simulation study. Reference line (in grey) indicates no bias.



Supplementary Figure 2: Power (significance level of α=10e-6) using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the case control simulation study.



Supplementary Figure 3: Observed coverage using the three strategies (NAIVE, DOSAGE and LC) according to minor allele frequency (MAF) and uncertainty (R^2), in the case control simulation study.



Supplementary Table 1: Case control simulation study results. NAIVE / DOSAGE / LC

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
5-15	5-15	5-15	5-15	1.0	1.02 / 1.09 / 1.09	99.1 / 96.1 / 100.0	0.0 / 0.0 / 0.0	41.6 / 19.8 / 27.6
5-15	25-35	5-15	5-15	1.0	1.02 / 1.03 / 1.03	97.0 / 95.2 / 99.7	0.0 / 0.0 / 0.0	9.6 / 2.1 / 5.2
5-15	45-55	5-15	5-15	1.0	1.02 / 1.00 / 1.00	96.5 / 96.2 / 99.0	0.0 / 0.0 / 0.0	2.8 / 0.4 / 1.1
5-15	65-75	5-15	5-15	1.0	1.02 / 1.02 / 1.02	97.1 / 95.8 / 99.1	0.0 / 0.0 / 0.0	2.3 / 0.6 / 0.7
5-15	85-95	5-15	5-15	1.0	1.00 / 1.00 / 0.99	96.5 / 93.8 / 96.3	0.0 / 0.0 / 0.0	1.7 / 0.1 / 0.3
5-15	25-35	5-15	25-35	1.0	0.99 / 1.00 / 1.00	94.9 / 95.1 / 98.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
5-15	45-55	5-15	25-35	1.0	1.00 / 1.00 / 1.00	93.6 / 94.9 / 95.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	65-75	5-15	25-35	1.0	1.00 / 1.01 / 1.01	95.6 / 94.9 / 96.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	25-35	1.0	1.00 / 1.00 / 1.00	94.6 / 95.5 / 96.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	1.0	1.01 / 1.01 / 1.01	94.5 / 94.6 / 95.5	0.0 / 0.0 / 0.0	0.1 / 0.1 / 0.1
5-15	65-75	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.3 / 94.9 / 95.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.7 / 94.4 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	1.0	1.00 / 1.01 / 1.01	94.9 / 95.5 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	1.0	1.00 / 1.00 / 1.00	96.1 / 95.7 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	1.0	0.99 / 0.99 / 0.99	94.0 / 93.8 / 93.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	1.0	1.02 / 1.01 / 1.02	95.1 / 95.4 / 99.7	0.0 / 0.0 / 0.0	0.4 / 0.0 / 1.8
25-35	25-35	5-15	5-15	1.0	0.99 / 0.98 / 0.98	94.5 / 95.8 / 97.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
25-35	45-55	5-15	5-15	1.0	1.00 / 0.99 / 0.99	95.6 / 93.6 / 96.0	0.0 / 0.0 / 0.0	0.2 / 0.1 / 0.1
25-35	65-75	5-15	5-15	1.0	0.98 / 0.98 / 0.98	94.7 / 94.8 / 96.0	0.0 / 0.0 / 0.0	0.2 / 0.0 / 0.0
25-35	85-95	5-15	5-15	1.0	1.01 / 1.00 / 1.00	93.7 / 94.4 / 95.4	0.0 / 0.0 / 0.0	0.1 / 0.0 / 0.0
25-35	25-35	5-15	25-35	1.0	0.99 / 0.99 / 0.99	95.3 / 94.4 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	5-15	25-35	1.0	1.00 / 0.99 / 0.99	95.2 / 95.4 / 96.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	1.0	1.00 / 1.00 / 1.00	95.0 / 96.0 / 96.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	1.0	1.01 / 1.01 / 1.01	94.3 / 95.2 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.2 / 94.7 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.1 / 95.9 / 96.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.8 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	1.0	1.00 / 0.99 / 0.99	95.0 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	1.0	1.00 / 1.00 / 1.00	95.8 / 95.8 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	1.0	1.00 / 1.00 / 1.00	94.8 / 94.9 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	1.0	0.98 / 1.00 / 1.01	96.0 / 95.5 / 99.8	0.0 / 0.0 / 0.0	4.7 / 1.4 / 3.5
45-55	25-35	5-15	5-15	1.0	0.99 / 1.00 / 1.00	95.1 / 92.8 / 97.4	0.0 / 0.0 / 0.0	0.2 / 0.0 / 0.1
45-55	45-55	5-15	5-15	1.0	1.02 / 1.01 / 1.01	94.8 / 93.9 / 96.6	0.0 / 0.0 / 0.0	0.1 / 0.0 / 0.0
45-55	65-75	5-15	5-15	1.0	1.00 / 0.99 / 0.99	94.6 / 95.0 / 96.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	5-15	1.0	1.00 / 0.99 / 0.99	93.9 / 95.1 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	25-35	5-15	25-35	1.0	1.01 / 1.01 / 1.01	95.1 / 94.7 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	5-15	25-35	1.0	1.00 / 0.99 / 0.99	94.7 / 95.9 / 96.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	5-15	25-35	1.0	1.00 / 1.00 / 1.00	95.1 / 94.3 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	1.0	1.00 / 1.00 / 1.00	94.2 / 94.5 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	1.0	1.00 / 1.00 / 1.00	94.9 / 94.1 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	1.0	1.00 / 1.00 / 1.00	94.2 / 94.5 / 94.9	0.1 / 0.1 / 0.1	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.0 / 96.0 / 96.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	1.0	1.00 / 1.00 / 1.00	95.0 / 95.1 / 95.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	1.0	1.00 / 1.00 / 1.00	94.0 / 94.1 / 94.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	1.0	1.00 / 1.00 / 1.00	95.0 / 94.2 / 94.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	1.0	1.00 / 0.98 / 0.98	94.5 / 95.3 / 97.6	0.0 / 0.0 / 0.0	1.2 / 1.2 / 1.4
25-35	25-35	25-35	5-15	1.0	1.00 / 1.00 / 1.00	94.9 / 94.9 / 95.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	5-15	1.0	1.00 / 1.01 / 1.01	94.3 / 95.3 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	5-15	1.0	1.00 / 1.01 / 1.01	94.8 / 94.5 / 94.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	1.0	1.00 / 1.00 / 1.00	95.6 / 94.8 / 95.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	1.0	1.00 / 0.99 / 0.99	95.7 / 95.0 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
25-35	45-55	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.8 / 94.9 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.5 / 94.7 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.9 / 94.3 / 94.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.3 / 95.7 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
25-35	65-75	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.3 / 95.1 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.0 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	1.0	1.00 / 0.99 / 0.99	95.2 / 95.6 / 95.6	0.0 / 0.0 / 0.0	0.1 / 0.1 / 0.3
25-35	85-95	25-35	65-75	1.0	1.00 / 1.00 / 1.00	95.6 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	1.0	1.00 / 1.00 / 1.00	94.5 / 94.9 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	1.0	1.00 / 1.00 / 1.00	94.9 / 95.5 / 98.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	25-35	25-35	5-15	1.0	1.00 / 1.00 / 1.00	95.7 / 94.3 / 95.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	5-15	1.0	1.00 / 1.00 / 1.00	95.3 / 94.8 / 95.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	5-15	1.0	1.00 / 1.00 / 1.00	94.8 / 95.5 / 95.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	1.0	1.00 / 1.01 / 1.01	95.3 / 95.3 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.8 / 95.8 / 96.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.7 / 95.8 / 96.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.6 / 94.7 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	1.0	1.00 / 1.00 / 1.00	95.4 / 95.0 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.0 / 94.8 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	1.0	1.00 / 1.00 / 1.00	94.6 / 95.3 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.3 / 96.0 / 96.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	1.0	1.00 / 1.00 / 1.00	94.4 / 95.5 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	1.0	1.00 / 1.00 / 1.00	95.2 / 95.0 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	1.0	1.00 / 1.00 / 1.00	94.4 / 94.2 / 94.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	5-15	45-55	5-15	1.0	1.01 / 0.99 / 0.99	96.1 / 94.6 / 99.2	0.0 / 0.0 / 0.0	2.9 / 2.9 / 5.4
45-55	25-35	45-55	5-15	1.0	1.00 / 1.00 / 1.00	94.7 / 93.8 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	5-15	1.0	1.00 / 1.00 / 1.00	95.1 / 94.7 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	5-15	1.0	1.00 / 0.99 / 0.99	95.7 / 94.7 / 94.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	5-15	1.0	0.99 / 0.99 / 0.99	94.7 / 94.8 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	1.0	1.00 / 1.00 / 1.00	94.8 / 95.6 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
45-55	45-55	45-55	25-35	1.0	1.00 / 1.00 / 1.00	96.1 / 95.8 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	1.0	1.00 / 1.00 / 1.00	96.1 / 96.3 / 96.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	1.0	1.00 / 0.99 / 0.99	94.6 / 94.9 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	1.0	1.00 / 1.00 / 1.00	95.7 / 95.9 / 95.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
45-55	65-75	45-55	45-55	1.0	1.00 / 1.00 / 1.00	95.4 / 95.3 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	1.0	1.00 / 1.00 / 1.00	94.0 / 94.6 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	1.0	1.00 / 1.00 / 1.00	94.3 / 93.3 / 93.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
45-55	85-95	45-55	65-75	1.0	1.00 / 1.00 / 1.00	95.2 / 95.3 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	1.0	1.00 / 1.00 / 1.00	96.4 / 95.4 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	1.5	1.09 / 1.41 / 1.34	97.1 / 97.4 / 99.8	0.0 / 0.0 / 0.0	44.9 / 20.6 / 29.9
5-15	25-35	5-15	5-15	1.5	1.13 / 1.49 / 1.56	92.9 / 95.4 / 99.1	0.0 / 0.0 / 0.0	10.3 / 2.3 / 4.7
5-15	45-55	5-15	5-15	1.5	1.18 / 1.46 / 1.50	90.0 / 94.9 / 98.0	0.0 / 0.0 / 0.0	3.1 / 0.2 / 1.4
5-15	65-75	5-15	5-15	1.5	1.19 / 1.52 / 1.57	91.6 / 95.0 / 98.2	0.0 / 0.0 / 0.0	2.5 / 0.3 / 0.7
5-15	85-95	5-15	5-15	1.5	1.24 / 1.48 / 1.53	92.0 / 95.2 / 96.9	0.0 / 0.0 / 0.0	2.4 / 0.0 / 0.1
5-15	25-35	5-15	25-35	1.5	1.18 / 1.48 / 1.53	86.3 / 95.1 / 98.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.7
5-15	45-55	5-15	25-35	1.5	1.22 / 1.48 / 1.51	82.4 / 95.9 / 97.0	0.1 / 0.1 / 0.0	0.0 / 0.0 / 0.0
5-15	65-75	5-15	25-35	1.5	1.24 / 1.51 / 1.54	83.7 / 95.2 / 95.9	0.1 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	25-35	1.5	1.28 / 1.49 / 1.51	85.4 / 95.1 / 95.4	0.0 / 0.2 / 0.1	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	1.5	1.25 / 1.50 / 1.52	79.0 / 95.1 / 95.8	0.2 / 0.4 / 0.0	0.0 / 0.0 / 0.0
5-15	65-75	5-15	45-55	1.5	1.28 / 1.49 / 1.50	82.0 / 94.9 / 95.3	0.1 / 0.2 / 0.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	45-55	1.5	1.33 / 1.50 / 1.52	85.8 / 93.6 / 93.9	0.5 / 1.5 / 0.9	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	1.5	1.32 / 1.49 / 1.50	83.9 / 95.6 / 95.8	0.9 / 1.5 / 1.1	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	1.5	1.38 / 1.50 / 1.51	90.6 / 95.9 / 95.7	1.9 / 2.6 / 2.3	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	1.5	1.43 / 1.49 / 1.49	93.6 / 95.9 / 96.1	6.1 / 6.3 / 6.0	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	1.5	1.09 / 1.56 / 1.63	81.1 / 95.3 / 99.3	0.0 / 0.0 / 0.0	0.4 / 0.4 / 2.7
25-35	25-35	5-15	5-15	1.5	1.11 / 1.50 / 1.56	79.6 / 95.7 / 97.6	0.0 / 0.0 / 0.0	0.3 / 0.0 / 0.2
25-35	45-55	5-15	5-15	1.5	1.17 / 1.50 / 1.55	81.9 / 95.8 / 96.7	0.0 / 0.1 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	5-15	1.5	1.21 / 1.49 / 1.53	82.3 / 94.8 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	5-15	1.5	1.25 / 1.50 / 1.54	84.5 / 95.9 / 95.8	0.0 / 0.3 / 0.2	0.1 / 0.0 / 0.0
25-35	25-35	5-15	25-35	1.5	1.16 / 1.50 / 1.54	56.7 / 95.3 / 96.1	0.1 / 0.3 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	5-15	25-35	1.5	1.21 / 1.49 / 1.52	61.4 / 96.2 / 96.4	0.1 / 0.5 / 0.1	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	1.5	1.24 / 1.48 / 1.50	65.3 / 95.3 / 95.5	0.4 / 1.3 / 0.6	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
25-35	85-95	5-15	25-35	1.5	1.27 / 1.48 / 1.50	68.5 / 95.0 / 95.0	1.0 / 2.9 / 1.9	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	1.5	1.23 / 1.49 / 1.51	47.7 / 95.5 / 95.6	0.8 / 2.5 / 1.4	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	1.5	1.27 / 1.49 / 1.51	58.7 / 94.9 / 94.7	4.5 / 6.3 / 5.1	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	1.5	1.32 / 1.49 / 1.51	71.7 / 94.3 / 94.8	6.7 / 12.7 / 10.8	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	1.5	1.31 / 1.48 / 1.49	67.5 / 94.6 / 94.6	9.2 / 13.7 / 11.8	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	1.5	1.37 / 1.49 / 1.50	80.6 / 94.8 / 95.0	22.2 / 27.5 / 25.2	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	1.5	1.44 / 1.50 / 1.50	92.6 / 95.1 / 95.1	48.7 / 50.9 / 50.3	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	1.5	1.09 / 1.50 / 1.52	89.4 / 95.0 / 99.6	0.0 / 0.0 / 0.0	4.9 / 0.8 / 4.5
45-55	25-35	5-15	5-15	1.5	1.11 / 1.42 / 1.48	78.8 / 93.4 / 96.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.4
45-55	45-55	5-15	5-15	1.5	1.17 / 1.48 / 1.53	78.4 / 94.5 / 96.0	0.0 / 0.0 / 0.0	0.1 / 0.0 / 0.1
45-55	65-75	5-15	5-15	1.5	1.20 / 1.47 / 1.52	78.8 / 95.5 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	5-15	1.5	1.22 / 1.46 / 1.50	82.4 / 94.8 / 95.0	0.0 / 0.3 / 0.0	0.1 / 0.0 / 0.0
45-55	25-35	5-15	25-35	1.5	1.16 / 1.49 / 1.53	51.8 / 95.3 / 96.4	0.0 / 0.2 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	5-15	25-35	1.5	1.19 / 1.48 / 1.50	52.6 / 94.5 / 95.0	0.4 / 0.8 / 0.4	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	1.5	1.23 / 1.47 / 1.50	60.5 / 94.7 / 95.5	0.8 / 1.4 / 0.8	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	1.5	1.27 / 1.48 / 1.51	68.4 / 95.2 / 95.8	1.4 / 4.4 / 2.9	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	1.5	1.23 / 1.48 / 1.50	39.4 / 94.7 / 94.6	1.8 / 4.4 / 2.9	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	1.5	1.27 / 1.49 / 1.51	52.5 / 94.6 / 94.7	6.0 / 12.6 / 10.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	1.5	1.31 / 1.48 / 1.50	63.9 / 95.0 / 94.7	12.2 / 18.7 / 17.1	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	1.5	1.32 / 1.49 / 1.50	61.6 / 94.5 / 94.7	15.0 / 21.3 / 19.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	1.5	1.37 / 1.49 / 1.50	78.6 / 95.4 / 95.6	31.5 / 38.5 / 37.1	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	1.5	1.43 / 1.49 / 1.50	89.4 / 95.1 / 95.2	61.9 / 65.4 / 65.1	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	1.5	1.05 / 1.51 / 1.58	6.6 / 95.0 / 96.9	0.0 / 0.0 / 0.0	0.8 / 0.8 / 1.4
25-35	25-35	25-35	5-15	1.5	1.08 / 1.48 / 1.53	2.2 / 96.2 / 97.1	0.0 / 0.2 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	5-15	1.5	1.11 / 1.49 / 1.53	2.3 / 96.3 / 96.4	0.1 / 0.5 / 0.2	0.0 / 0.0 / 0.0
25-35	65-75	25-35	5-15	1.5	1.13 / 1.48 / 1.51	3.1 / 94.8 / 95.6	0.1 / 1.1 / 0.6	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	1.5	1.14 / 1.48 / 1.50	2.5 / 96.1 / 96.5	0.4 / 1.9 / 1.0	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	1.5	1.13 / 1.47 / 1.50	3.2 / 95.6 / 95.8	0.3 / 1.2 / 0.2	0.0 / 0.0 / 0.1
25-35	45-55	25-35	25-35	1.5	1.17 / 1.48 / 1.51	4.3 / 95.5 / 95.4	2.1 / 6.8 / 4.9	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	1.5	1.20 / 1.49 / 1.51	7.6 / 95.0 / 94.9	7.5 / 18.4 / 14.9	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	1.5	1.24 / 1.49 / 1.51	13.0 / 94.8 / 95.0	16.3 / 32.4 / 28.9	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	1.5	1.22 / 1.49 / 1.51	7.2 / 94.1 / 94.4	12.4 / 27.2 / 23.5	0.0 / 0.0 / 0.3
25-35	65-75	25-35	45-55	1.5	1.26 / 1.48 / 1.50	15.8 / 94.9 / 94.8	30.9 / 50.5 / 47.1	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	1.5	1.30 / 1.49 / 1.50	30.8 / 94.6 / 94.9	59.3 / 75.1 / 73.9	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	1.5	1.31 / 1.49 / 1.51	34.4 / 95.5 / 95.3	67.6 / 80.3 / 78.7	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	1.5	1.37 / 1.49 / 1.50	62.3 / 94.6 / 95.1	90.0 / 93.9 / 93.5	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	1.5	1.43 / 1.50 / 1.50	85.7 / 94.2 / 94.2	98.5 / 98.7 / 98.6	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	1.5	1.06 / 1.48 / 1.57	31.6 / 96.0 / 98.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	25-35	25-35	5-15	1.5	1.08 / 1.45 / 1.50	1.5 / 94.6 / 96.0	0.0 / 0.1 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	5-15	1.5	1.10 / 1.47 / 1.52	0.5 / 94.7 / 95.7	0.1 / 0.8 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	5-15	1.5	1.12 / 1.47 / 1.51	1.4 / 96.0 / 96.6	0.6 / 1.4 / 0.9	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	1.5	1.14 / 1.47 / 1.51	1.5 / 95.6 / 95.6	1.3 / 4.0 / 2.3	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	1.5	1.14 / 1.48 / 1.51	1.9 / 95.0 / 95.1	0.7 / 2.0 / 0.9	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	1.5	1.17 / 1.48 / 1.51	1.1 / 94.8 / 95.7	3.0 / 11.1 / 7.3	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	1.5	1.20 / 1.47 / 1.50	4.0 / 93.9 / 94.2	11.2 / 27.4 / 24.6	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	1.5	1.23 / 1.48 / 1.50	7.9 / 96.0 / 95.3	24.1 / 45.3 / 42.6	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	1.5	1.21 / 1.48 / 1.50	2.6 / 94.3 / 95.4	16.9 / 35.5 / 32.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	1.5	1.26 / 1.49 / 1.51	9.8 / 94.0 / 94.4	48.8 / 68.0 / 66.3	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	1.5	1.30 / 1.49 / 1.50	23.3 / 95.9 / 95.6	72.1 / 86.7 / 86.3	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	1.5	1.31 / 1.49 / 1.50	24.6 / 95.9 / 96.1	80.9 / 91.3 / 91.1	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	1.5	1.37 / 1.49 / 1.50	53.1 / 95.2 / 95.5	95.5 / 98.0 / 98.1	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	1.5	1.44 / 1.50 / 1.51	85.0 / 95.3 / 95.2	99.7 / 99.8 / 99.8	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	1.5	1.07 / 1.45 / 1.53	57.6 / 94.7 / 98.3	0.0 / 0.1 / 0.0	3.6 / 3.6 / 6.0
45-55	25-35	45-55	5-15	1.5	1.09 / 1.44 / 1.49	25.0 / 93.8 / 95.2	0.1 / 0.2 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	5-15	1.5	1.12 / 1.48 / 1.52	21.7 / 94.1 / 94.6	0.1 / 1.1 / 0.1	0.0 / 0.0 / 0.0
45-55	65-75	45-55	5-15	1.5	1.15 / 1.49 / 1.53	22.7 / 94.9 / 95.1	0.4 / 2.5 / 1.3	0.0 / 0.0 / 0.0
45-55	85-95	45-55	5-15	1.5	1.18 / 1.48 / 1.52	30.8 / 95.3 / 95.4	1.3 / 4.8 / 2.4	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	1.5	1.14 / 1.48 / 1.52	1.8 / 94.7 / 95.5	0.7 / 3.7 / 1.9	0.0 / 0.0 / 0.0
45-55	45-55	45-55	25-35	1.5	1.17 / 1.47 / 1.51	1.5 / 93.9 / 94.7	5.1 / 15.3 / 11.5	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	1.5	1.20 / 1.47 / 1.50	2.5 / 94.3 / 94.5	14.2 / 35.0 / 30.7	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	1.5	1.24 / 1.47 / 1.49	6.3 / 94.2 / 95.5	32.2 / 55.1 / 52.2	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	1.5	1.21 / 1.47 / 1.50	0.9 / 95.3 / 95.9	25.1 / 47.0 / 43.6	0.2 / 0.2 / 0.5
45-55	65-75	45-55	45-55	1.5	1.25 / 1.48 / 1.50	3.9 / 94.0 / 94.2	58.1 / 78.0 / 77.5	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	1.5	1.30 / 1.48 / 1.50	16.3 / 95.3 / 95.8	84.5 / 94.0 / 93.5	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	1.5	1.31 / 1.48 / 1.50	17.0 / 94.9 / 95.6	91.1 / 97.2 / 97.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	1.5	1.37 / 1.49 / 1.50	46.4 / 94.0 / 93.9	98.9 / 99.7 / 99.7	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	1.5	1.44 / 1.49 / 1.50	82.9 / 94.4 / 94.3	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	2.0	1.14 / 1.94 / 1.75	94.1 / 98.3 / 100.0	0.0 / 0.0 / 0.0	45.7 / 21.4 / 32.9
5-15	25-35	5-15	5-15	2.0	1.22 / 1.92 / 1.99	86.8 / 95.8 / 98.7	0.0 / 0.0 / 0.0	10.0 / 2.5 / 7.9
5-15	45-55	5-15	5-15	2.0	1.29 / 1.90 / 2.04	79.0 / 94.4 / 96.8	0.0 / 0.0 / 0.0	4.4 / 0.8 / 3.1
5-15	65-75	5-15	5-15	2.0	1.35 / 1.88 / 2.01	82.0 / 94.9 / 96.8	0.0 / 0.0 / 0.0	3.4 / 0.5 / 1.9
5-15	85-95	5-15	5-15	2.0	1.46 / 1.94 / 2.08	81.6 / 94.8 / 96.6	0.0 / 0.7 / 0.1	2.2 / 0.2 / 0.9
5-15	25-35	5-15	25-35	2.0	1.35 / 1.99 / 2.13	70.9 / 94.9 / 97.5	0.0 / 0.3 / 0.0	0.1 / 0.0 / 1.2
5-15	45-55	5-15	25-35	2.0	1.36 / 1.87 / 1.97	58.9 / 95.4 / 96.3	0.0 / 0.5 / 0.0	0.0 / 0.0 / 0.1
5-15	65-75	5-15	25-35	2.0	1.45 / 1.97 / 2.08	63.0 / 96.0 / 97.2	0.1 / 1.4 / 0.4	0.0 / 0.0 / 0.0
5-15	85-95	5-15	25-35	2.0	1.52 / 1.94 / 2.02	69.1 / 94.5 / 95.3	1.6 / 4.6 / 2.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
5-15	45-55	5-15	45-55	2.0	1.43 / 1.96 / 2.05	49.6 / 94.1 / 95.7	1.6 / 4.3 / 1.6	0.0 / 0.0 / 0.0
5-15	65-75	5-15	45-55	2.0	1.50 / 1.95 / 2.02	57.1 / 94.7 / 95.3	4.6 / 8.9 / 4.5	0.0 / 0.0 / 0.0
5-15	85-95	5-15	45-55	2.0	1.59 / 1.95 / 2.00	62.4 / 95.0 / 95.9	10.6 / 18.7 / 14.1	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	2.0	1.60 / 1.96 / 2.01	63.1 / 94.7 / 94.7	12.8 / 17.8 / 13.4	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	2.0	1.72 / 1.98 / 2.01	77.9 / 95.0 / 95.1	29.4 / 36.5 / 31.7	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	2.0	1.87 / 1.99 / 2.01	91.0 / 95.6 / 95.5	58.7 / 62.7 / 61.3	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	2.0	1.15 / 1.80 / 1.90	57.5 / 94.3 / 98.1	0.0 / 0.0 / 0.0	0.4 / 0.6 / 5.4
25-35	25-35	5-15	5-15	2.0	1.23 / 1.89 / 2.12	56.2 / 94.3 / 96.2	0.0 / 0.1 / 0.0	0.0 / 0.0 / 0.7
25-35	45-55	5-15	5-15	2.0	1.28 / 1.85 / 2.05	58.8 / 93.9 / 95.4	0.0 / 0.7 / 0.1	0.0 / 0.0 / 0.0
25-35	65-75	5-15	5-15	2.0	1.36 / 1.87 / 2.01	60.8 / 94.7 / 96.5	0.2 / 1.1 / 0.2	0.0 / 0.0 / 0.2
25-35	85-95	5-15	5-15	2.0	1.41 / 1.91 / 2.08	67.7 / 94.0 / 94.5	0.1 / 3.3 / 0.8	0.1 / 0.0 / 0.0
25-35	25-35	5-15	25-35	2.0	1.27 / 1.87 / 2.02	16.7 / 94.2 / 96.0	0.1 / 1.4 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	5-15	25-35	2.0	1.35 / 1.88 / 2.00	21.3 / 94.5 / 95.5	1.8 / 7.4 / 2.4	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	2.0	1.42 / 1.92 / 2.04	29.6 / 95.5 / 96.3	4.5 / 17.4 / 10.7	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	2.0	1.49 / 1.91 / 2.01	38.1 / 95.2 / 96.4	10.2 / 31.8 / 23.9	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	2.0	1.40 / 1.92 / 2.01	11.1 / 94.4 / 95.5	13.1 / 33.0 / 22.7	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	2.0	1.49 / 1.94 / 2.02	22.8 / 93.9 / 94.8	36.8 / 55.7 / 47.9	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	2.0	1.58 / 1.94 / 2.01	34.9 / 93.8 / 94.0	57.5 / 75.5 / 71.2	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	2.0	1.59 / 1.96 / 2.01	31.6 / 94.8 / 95.2	64.5 / 75.5 / 71.9	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	2.0	1.69 / 1.95 / 1.99	53.7 / 94.6 / 94.9	85.0 / 89.7 / 88.2	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	2.0	1.85 / 1.98 / 2.00	83.5 / 95.1 / 95.7	97.1 / 98.0 / 97.6	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	2.0	1.14 / 1.77 / 1.87	74.9 / 96.6 / 99.0	0.0 / 0.0 / 0.0	6.0 / 1.7 / 7.3
45-55	25-35	5-15	5-15	2.0	1.23 / 1.84 / 2.06	53.7 / 95.7 / 97.8	0.0 / 0.1 / 0.0	0.3 / 0.0 / 0.5
45-55	45-55	5-15	5-15	2.0	1.27 / 1.82 / 2.02	51.9 / 94.0 / 96.8	0.0 / 0.4 / 0.0	0.0 / 0.0 / 0.3
45-55	65-75	5-15	5-15	2.0	1.34 / 1.85 / 2.04	55.4 / 93.5 / 96.0	0.0 / 2.2 / 0.1	0.0 / 0.0 / 0.1
45-55	85-95	5-15	5-15	2.0	1.42 / 1.87 / 2.06	63.2 / 94.6 / 96.2	0.1 / 4.5 / 1.0	0.0 / 0.0 / 0.0
45-55	25-35	5-15	25-35	2.0	1.26 / 1.87 / 2.04	14.6 / 94.4 / 95.5	0.1 / 2.3 / 0.2	0.0 / 0.0 / 0.0
45-55	45-55	5-15	25-35	2.0	1.32 / 1.87 / 2.01	16.0 / 93.3 / 95.2	2.0 / 9.8 / 3.7	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	2.0	1.40 / 1.87 / 2.00	19.8 / 93.1 / 95.7	7.1 / 22.7 / 15.3	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	2.0	1.47 / 1.89 / 2.01	33.7 / 93.4 / 95.6	14.0 / 39.0 / 32.9	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	2.0	1.39 / 1.91 / 2.02	8.4 / 93.6 / 94.5	18.7 / 40.4 / 30.2	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	2.0	1.48 / 1.92 / 2.01	13.7 / 94.1 / 96.2	45.8 / 65.7 / 60.9	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	2.0	1.57 / 1.92 / 2.00	26.7 / 92.2 / 94.2	64.5 / 81.1 / 78.1	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	2.0	1.58 / 1.94 / 2.01	25.7 / 95.0 / 95.9	76.4 / 86.7 / 84.1	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	2.0	1.69 / 1.95 / 2.00	48.6 / 92.5 / 93.4	89.1 / 93.1 / 92.2	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	2.0	1.85 / 1.99 / 2.01	82.3 / 94.3 / 94.9	99.1 / 99.5 / 99.5	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	2.0	1.08 / 1.90 / 2.15	0.1 / 94.7 / 97.1	0.0 / 0.2 / 0.0	0.8 / 0.9 / 1.9
25-35	25-35	25-35	5-15	2.0	1.13 / 1.83 / 2.01	0.0 / 92.7 / 93.7	0.2 / 1.5 / 0.1	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
25-35	45-55	25-35	5-15	2.0	1.17 / 1.86 / 2.03	0.0 / 94.4 / 96.1	0.9 / 6.2 / 2.8	0.0 / 0.0 / 0.0
25-35	65-75	25-35	5-15	2.0	1.20 / 1.89 / 2.02	0.0 / 94.3 / 96.2	3.6 / 15.8 / 12.5	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	2.0	1.24 / 1.88 / 2.02	0.0 / 93.1 / 95.1	11.2 / 30.6 / 27.6	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	2.0	1.22 / 1.89 / 2.04	0.0 / 94.7 / 97.1	5.4 / 21.9 / 13.2	0.0 / 0.0 / 0.1
25-35	45-55	25-35	25-35	2.0	1.28 / 1.87 / 1.99	0.0 / 91.4 / 96.1	26.0 / 56.0 / 50.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	2.0	1.34 / 1.89 / 2.00	0.0 / 91.8 / 94.5	56.3 / 84.1 / 80.6	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	2.0	1.42 / 1.91 / 2.01	0.2 / 92.8 / 94.8	84.2 / 96.8 / 96.6	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	2.0	1.36 / 1.89 / 2.00	0.1 / 91.6 / 95.6	73.5 / 93.6 / 91.9	0.0 / 0.0 / 0.1
25-35	65-75	25-35	45-55	2.0	1.45 / 1.92 / 2.00	0.6 / 92.3 / 94.8	97.0 / 99.4 / 99.4	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	2.0	1.55 / 1.94 / 2.00	1.5 / 93.4 / 96.9	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	2.0	1.56 / 1.94 / 2.01	2.2 / 93.4 / 96.6	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	2.0	1.68 / 1.95 / 2.00	16.9 / 93.5 / 94.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	2.0	1.84 / 1.98 / 2.00	68.3 / 95.3 / 95.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	2.0	1.09 / 1.80 / 2.05	2.7 / 92.6 / 96.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 1.2
45-55	25-35	25-35	5-15	2.0	1.13 / 1.82 / 2.04	0.0 / 92.9 / 95.7	0.2 / 2.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	5-15	2.0	1.16 / 1.83 / 2.02	0.0 / 92.9 / 95.0	2.1 / 8.2 / 3.1	0.0 / 0.0 / 0.0
45-55	65-75	25-35	5-15	2.0	1.20 / 1.85 / 2.03	0.0 / 91.8 / 95.3	6.2 / 22.1 / 16.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	2.0	1.22 / 1.84 / 2.00	0.0 / 91.7 / 96.1	11.6 / 35.8 / 32.2	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	2.0	1.22 / 1.82 / 1.99	0.0 / 92.6 / 95.4	5.0 / 26.0 / 16.2	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	2.0	1.28 / 1.86 / 2.01	0.0 / 90.3 / 95.4	35.2 / 69.8 / 65.1	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	2.0	1.33 / 1.86 / 2.00	0.0 / 89.1 / 95.4	67.4 / 92.3 / 91.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	2.0	1.39 / 1.87 / 1.99	0.0 / 88.5 / 95.1	87.4 / 98.4 / 98.5	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	2.0	1.36 / 1.88 / 2.01	0.1 / 89.7 / 94.1	84.4 / 96.5 / 95.5	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	2.0	1.44 / 1.90 / 2.00	0.0 / 89.5 / 95.1	98.5 / 99.9 / 99.9	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	2.0	1.53 / 1.92 / 2.01	0.5 / 91.7 / 95.7	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	2.0	1.57 / 1.94 / 2.01	1.1 / 93.1 / 95.0	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	2.0	1.68 / 1.95 / 2.00	11.2 / 93.0 / 94.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	2.0	1.84 / 1.98 / 2.00	65.6 / 95.6 / 96.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	2.0	1.12 / 1.75 / 2.01	23.3 / 95.1 / 97.8	0.1 / 0.2 / 0.0	3.4 / 3.4 / 8.0
45-55	25-35	45-55	5-15	2.0	1.15 / 1.79 / 2.03	1.1 / 92.5 / 95.2	0.1 / 1.8 / 0.1	0.0 / 0.0 / 0.0
45-55	45-55	45-55	5-15	2.0	1.19 / 1.82 / 2.04	1.0 / 91.4 / 96.0	0.3 / 7.2 / 0.8	0.0 / 0.0 / 0.0
45-55	65-75	45-55	5-15	2.0	1.23 / 1.83 / 2.01	0.8 / 92.8 / 96.7	1.7 / 16.8 / 9.8	0.0 / 0.0 / 0.0
45-55	85-95	45-55	5-15	2.0	1.27 / 1.86 / 2.02	1.4 / 92.9 / 96.4	6.2 / 35.1 / 28.8	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	2.0	1.22 / 1.82 / 2.03	0.0 / 90.7 / 94.6	9.9 / 34.8 / 25.8	0.0 / 0.0 / 0.5
45-55	45-55	45-55	25-35	2.0	1.28 / 1.84 / 2.01	0.0 / 89.2 / 94.8	41.1 / 78.4 / 75.4	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	2.0	1.34 / 1.86 / 2.01	0.0 / 88.5 / 95.1	76.0 / 95.8 / 95.4	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	2.0	1.41 / 1.89 / 2.01	0.1 / 90.2 / 95.4	94.9 / 99.9 / 99.8	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	2.0	1.35 / 1.86 / 2.00	0.0 / 88.8 / 95.4	90.2 / 98.5 / 98.4	0.1 / 0.1 / 0.6

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	45-55	45-55	2.0	1.44 / 1.88 / 1.99	0.0 / 87.0 / 94.9	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	2.0	1.53 / 1.90 / 2.00	0.1 / 88.9 / 94.9	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	2.0	1.55 / 1.91 / 2.00	0.3 / 89.0 / 95.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.2
45-55	85-95	45-55	65-75	2.0	1.68 / 1.95 / 2.01	8.0 / 93.5 / 95.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	2.0	1.84 / 1.98 / 2.00	59.3 / 95.0 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	2.5	1.14 / 1.99 / 1.77	87.3 / 96.2 / 98.5	0.0 / 0.0 / 0.0	42.3 / 21.6 / 35.5
5-15	25-35	5-15	5-15	2.5	1.35 / 2.22 / 2.27	78.4 / 94.4 / 96.7	0.0 / 0.0 / 0.0	10.8 / 2.8 / 11.9
5-15	45-55	5-15	5-15	2.5	1.42 / 2.32 / 2.57	68.8 / 94.6 / 96.2	0.1 / 0.2 / 0.0	3.7 / 0.3 / 4.6
5-15	65-75	5-15	5-15	2.5	1.47 / 2.31 / 2.54	70.5 / 93.3 / 95.5	0.0 / 0.2 / 0.0	2.9 / 0.2 / 2.8
5-15	85-95	5-15	5-15	2.5	1.61 / 2.29 / 2.59	73.3 / 92.9 / 95.3	0.1 / 1.9 / 0.1	2.2 / 0.4 / 1.4
5-15	25-35	5-15	25-35	2.5	1.43 / 2.30 / 2.59	53.7 / 94.8 / 96.5	0.0 / 0.4 / 0.0	0.0 / 0.0 / 1.9
5-15	45-55	5-15	25-35	2.5	1.48 / 2.27 / 2.52	41.2 / 94.2 / 95.6	0.3 / 2.2 / 0.1	0.0 / 0.0 / 0.2
5-15	65-75	5-15	25-35	2.5	1.59 / 2.35 / 2.57	45.3 / 95.9 / 97.1	1.2 / 5.5 / 0.9	0.0 / 0.0 / 0.2
5-15	85-95	5-15	25-35	2.5	1.71 / 2.36 / 2.55	50.9 / 94.0 / 95.6	5.2 / 14.5 / 5.9	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	2.5	1.58 / 2.37 / 2.59	28.0 / 92.9 / 94.6	3.4 / 13.2 / 3.9	0.0 / 0.0 / 0.0
5-15	65-75	5-15	45-55	2.5	1.71 / 2.40 / 2.57	39.0 / 94.4 / 95.0	13.6 / 25.7 / 13.9	0.0 / 0.0 / 0.0
5-15	85-95	5-15	45-55	2.5	1.83 / 2.38 / 2.51	46.2 / 93.4 / 94.8	32.5 / 47.6 / 36.3	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	2.5	1.86 / 2.45 / 2.57	49.8 / 95.5 / 95.2	35.0 / 45.7 / 34.9	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	2.5	2.02 / 2.43 / 2.51	65.9 / 93.6 / 94.8	59.7 / 66.5 / 62.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	2.5	2.28 / 2.49 / 2.53	88.8 / 94.3 / 94.3	85.4 / 88.3 / 87.1	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	2.5	1.20 / 2.12 / 2.30	40.4 / 94.9 / 97.5	0.1 / 0.2 / 0.0	0.4 / 0.9 / 8.6
25-35	25-35	5-15	5-15	2.5	1.30 / 2.20 / 2.65	40.0 / 95.3 / 97.2	0.1 / 0.3 / 0.0	0.3 / 0.0 / 1.5
25-35	45-55	5-15	5-15	2.5	1.39 / 2.17 / 2.57	41.1 / 93.3 / 94.7	0.1 / 1.8 / 0.1	0.2 / 0.0 / 0.4
25-35	65-75	5-15	5-15	2.5	1.49 / 2.25 / 2.64	47.6 / 93.1 / 95.1	0.3 / 4.3 / 1.2	0.2 / 0.0 / 0.2
25-35	85-95	5-15	5-15	2.5	1.54 / 2.24 / 2.58	49.5 / 92.9 / 94.9	0.6 / 11.0 / 5.6	0.2 / 0.0 / 0.1
25-35	25-35	5-15	25-35	2.5	1.34 / 2.24 / 2.58	5.7 / 92.2 / 95.3	1.0 / 6.1 / 0.7	0.0 / 0.0 / 0.0
25-35	45-55	5-15	25-35	2.5	1.46 / 2.26 / 2.56	8.4 / 91.9 / 95.1	6.2 / 22.2 / 11.1	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	2.5	1.56 / 2.28 / 2.54	11.8 / 90.7 / 94.3	17.4 / 46.0 / 31.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	2.5	1.67 / 2.29 / 2.52	21.3 / 90.6 / 94.4	30.0 / 63.0 / 52.7	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	2.5	1.53 / 2.29 / 2.51	2.8 / 91.6 / 95.1	35.0 / 63.9 / 51.3	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	2.5	1.67 / 2.34 / 2.54	7.2 / 91.8 / 94.8	68.2 / 84.7 / 79.7	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	2.5	1.80 / 2.35 / 2.52	16.9 / 92.0 / 95.2	84.6 / 94.8 / 92.2	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	2.5	1.81 / 2.39 / 2.54	13.9 / 92.6 / 94.1	91.0 / 96.1 / 94.2	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	2.5	1.99 / 2.40 / 2.51	36.4 / 92.8 / 95.4	98.5 / 99.3 / 99.2	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	2.5	2.24 / 2.46 / 2.51	75.3 / 93.8 / 93.7	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	2.5	1.18 / 2.02 / 2.16	60.0 / 94.4 / 97.8	0.0 / 0.0 / 0.0	5.9 / 1.8 / 10.9
45-55	25-35	5-15	5-15	2.5	1.26 / 2.07 / 2.50	35.4 / 92.5 / 96.6	0.0 / 0.6 / 0.0	0.2 / 0.0 / 2.1
45-55	45-55	5-15	5-15	2.5	1.36 / 2.11 / 2.59	36.2 / 92.2 / 95.4	0.1 / 1.5 / 0.1	0.0 / 0.0 / 0.5

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	5-15	5-15	2.5	1.42 / 2.13 / 2.57	36.3 / 91.0 / 95.1	0.9 / 4.7 / 1.6	0.0 / 0.0 / 0.0
45-55	85-95	5-15	5-15	2.5	1.52 / 2.14 / 2.59	40.8 / 90.5 / 95.7	1.2 / 11.6 / 7.1	0.0 / 0.0 / 0.0
45-55	25-35	5-15	25-35	2.5	1.34 / 2.16 / 2.52	3.4 / 93.2 / 96.1	1.8 / 7.7 / 0.9	0.0 / 0.0 / 0.1
45-55	45-55	5-15	25-35	2.5	1.43 / 2.21 / 2.57	5.1 / 90.7 / 95.2	6.0 / 30.3 / 13.9	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	2.5	1.53 / 2.21 / 2.53	6.4 / 89.0 / 95.2	20.3 / 51.6 / 41.2	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	2.5	1.63 / 2.22 / 2.52	11.7 / 87.7 / 96.1	36.1 / 70.5 / 63.4	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	2.5	1.53 / 2.29 / 2.56	2.0 / 89.8 / 94.5	45.9 / 71.9 / 64.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	2.5	1.64 / 2.29 / 2.53	4.3 / 88.0 / 94.0	74.1 / 91.7 / 89.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	2.5	1.77 / 2.30 / 2.52	9.8 / 87.1 / 95.3	90.3 / 97.0 / 96.3	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	2.5	1.80 / 2.34 / 2.50	9.6 / 91.5 / 95.1	94.8 / 98.2 / 97.3	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	2.5	1.96 / 2.37 / 2.51	27.3 / 91.6 / 95.6	99.3 / 99.8 / 99.6	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	2.5	2.23 / 2.45 / 2.50	72.1 / 94.1 / 95.8	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	2.5	1.10 / 2.17 / 2.66	0.1 / 93.9 / 96.4	0.0 / 0.2 / 0.0	1.2 / 1.2 / 2.3
25-35	25-35	25-35	5-15	2.5	1.17 / 2.16 / 2.58	0.0 / 92.7 / 96.3	1.0 / 4.2 / 0.7	0.0 / 0.0 / 0.0
25-35	45-55	25-35	5-15	2.5	1.21 / 2.18 / 2.55	0.0 / 91.4 / 94.9	4.7 / 19.5 / 12.7	0.0 / 0.0 / 0.0
25-35	65-75	25-35	5-15	2.5	1.26 / 2.19 / 2.51	0.0 / 88.8 / 95.0	11.7 / 37.3 / 34.4	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	2.5	1.30 / 2.22 / 2.50	0.0 / 90.1 / 95.3	24.8 / 60.5 / 62.7	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	2.5	1.28 / 2.16 / 2.50	0.0 / 88.5 / 95.6	16.7 / 50.1 / 38.7	0.0 / 0.0 / 0.1
25-35	45-55	25-35	25-35	2.5	1.36 / 2.21 / 2.51	0.0 / 86.2 / 95.4	57.6 / 91.1 / 89.3	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	2.5	1.45 / 2.25 / 2.50	0.0 / 85.5 / 96.0	89.7 / 98.8 / 98.7	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	2.5	1.55 / 2.28 / 2.50	0.0 / 85.2 / 94.5	98.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	2.5	1.49 / 2.27 / 2.53	0.0 / 83.6 / 95.1	96.9 / 99.8 / 99.7	0.0 / 0.0 / 0.3
25-35	65-75	25-35	45-55	2.5	1.61 / 2.30 / 2.50	0.0 / 85.2 / 94.8	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	2.5	1.75 / 2.34 / 2.50	0.2 / 85.7 / 94.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	2.5	1.78 / 2.36 / 2.51	0.0 / 89.1 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.1
25-35	85-95	25-35	65-75	2.5	1.97 / 2.40 / 2.50	4.5 / 89.8 / 94.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	2.5	2.24 / 2.46 / 2.51	51.2 / 95.3 / 95.3	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	2.5	1.11 / 2.07 / 2.56	0.1 / 93.3 / 97.3	0.0 / 0.5 / 0.0	0.0 / 0.0 / 3.6
45-55	25-35	25-35	5-15	2.5	1.16 / 2.07 / 2.55	0.0 / 88.5 / 95.2	0.7 / 6.7 / 1.2	0.0 / 0.0 / 0.0
45-55	45-55	25-35	5-15	2.5	1.20 / 2.13 / 2.57	0.0 / 89.4 / 96.7	4.9 / 24.6 / 15.4	0.0 / 0.0 / 0.0
45-55	65-75	25-35	5-15	2.5	1.24 / 2.14 / 2.55	0.0 / 86.4 / 95.0	16.2 / 46.8 / 44.7	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	2.5	1.28 / 2.12 / 2.50	0.0 / 82.3 / 95.2	28.0 / 65.8 / 70.5	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	2.5	1.28 / 2.14 / 2.52	0.0 / 84.0 / 95.6	21.8 / 60.2 / 51.7	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	2.5	1.35 / 2.16 / 2.51	0.0 / 79.1 / 94.9	64.6 / 93.6 / 92.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	2.5	1.43 / 2.20 / 2.52	0.0 / 79.3 / 95.5	95.3 / 99.6 / 99.6	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	2.5	1.52 / 2.22 / 2.51	0.0 / 76.1 / 94.6	99.1 / 99.8 / 99.9	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	2.5	1.47 / 2.22 / 2.50	0.0 / 77.2 / 94.6	97.8 / 100.0 / 99.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	2.5	1.59 / 2.26 / 2.51	0.0 / 78.4 / 96.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
45-55	85-95	25-35	45-55	2.5	1.71 / 2.29 / 2.50	0.1 / 79.7 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	2.5	1.76 / 2.33 / 2.51	0.0 / 85.0 / 95.7	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	2.5	1.95 / 2.37 / 2.51	1.3 / 86.5 / 95.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	2.5	2.22 / 2.45 / 2.50	44.5 / 93.7 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	2.5	1.13 / 2.03 / 2.50	7.3 / 91.8 / 97.0	0.1 / 0.0 / 0.0	4.0 / 4.0 / 14.7
45-55	25-35	45-55	5-15	2.5	1.18 / 2.03 / 2.59	0.0 / 89.4 / 95.9	0.1 / 4.2 / 0.1	0.0 / 0.0 / 0.1
45-55	45-55	45-55	5-15	2.5	1.23 / 2.05 / 2.51	0.0 / 83.6 / 97.1	1.5 / 16.8 / 8.5	0.0 / 0.0 / 0.0
45-55	65-75	45-55	5-15	2.5	1.28 / 2.10 / 2.51	0.0 / 83.0 / 95.4	6.5 / 40.9 / 34.3	0.0 / 0.0 / 0.0
45-55	85-95	45-55	5-15	2.5	1.35 / 2.14 / 2.52	0.1 / 83.8 / 94.6	12.9 / 57.5 / 63.5	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	2.5	1.27 / 2.05 / 2.50	0.0 / 74.6 / 95.3	20.5 / 64.3 / 52.9	0.0 / 0.0 / 0.3
45-55	45-55	45-55	25-35	2.5	1.35 / 2.12 / 2.52	0.0 / 71.0 / 94.4	68.5 / 95.7 / 95.5	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	2.5	1.44 / 2.16 / 2.50	0.0 / 70.1 / 96.3	95.9 / 99.9 / 99.9	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	2.5	1.54 / 2.22 / 2.51	0.0 / 74.5 / 95.6	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	2.5	1.45 / 2.17 / 2.52	0.0 / 69.2 / 94.9	99.7 / 100.0 / 100.0	0.0 / 0.0 / 0.2
45-55	65-75	45-55	45-55	2.5	1.57 / 2.22 / 2.50	0.0 / 69.3 / 94.7	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	2.5	1.71 / 2.27 / 2.50	0.0 / 74.9 / 95.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	2.5	1.74 / 2.29 / 2.50	0.0 / 76.0 / 95.7	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	2.5	1.94 / 2.36 / 2.51	0.4 / 84.6 / 96.1	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	2.5	2.20 / 2.43 / 2.50	35.4 / 92.2 / 94.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	3.0	1.18 / 2.32 / 1.99	82.9 / 96.3 / 98.4	0.0 / 0.0 / 0.0	45.5 / 21.8 / 38.2
5-15	25-35	5-15	5-15	3.0	1.32 / 2.61 / 2.68	69.3 / 95.4 / 96.2	0.0 / 0.0 / 0.0	9.2 / 2.6 / 15.9
5-15	45-55	5-15	5-15	3.0	1.47 / 2.58 / 2.93	62.1 / 94.2 / 95.1	0.2 / 0.4 / 0.0	3.2 / 0.6 / 7.1
5-15	65-75	5-15	5-15	3.0	1.63 / 2.67 / 3.05	63.5 / 95.5 / 97.5	0.0 / 0.5 / 0.0	3.2 / 0.3 / 5.4
5-15	85-95	5-15	5-15	3.0	1.67 / 2.60 / 3.00	65.2 / 94.7 / 97.1	0.1 / 2.9 / 0.1	2.5 / 0.0 / 2.6
5-15	25-35	5-15	25-35	3.0	1.47 / 2.56 / 2.96	37.7 / 93.5 / 95.4	0.2 / 0.6 / 0.0	0.0 / 0.1 / 3.3
5-15	45-55	5-15	25-35	3.0	1.58 / 2.66 / 3.08	27.8 / 94.3 / 96.1	0.6 / 5.5 / 0.0	0.0 / 0.0 / 0.9
5-15	65-75	5-15	25-35	3.0	1.73 / 2.74 / 3.15	32.9 / 93.7 / 96.3	4.8 / 13.6 / 3.2	0.0 / 0.0 / 0.8
5-15	85-95	5-15	25-35	3.0	1.89 / 2.78 / 3.15	40.1 / 93.5 / 96.4	8.7 / 27.5 / 10.7	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	3.0	1.72 / 2.71 / 3.09	20.3 / 93.6 / 95.6	8.2 / 25.9 / 7.3	0.1 / 0.1 / 0.3
5-15	65-75	5-15	45-55	3.0	1.86 / 2.79 / 3.09	24.0 / 93.3 / 96.6	25.1 / 43.3 / 26.9	0.0 / 0.0 / 0.2
5-15	85-95	5-15	45-55	3.0	2.05 / 2.78 / 3.02	34.7 / 92.6 / 95.1	49.6 / 64.8 / 54.1	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	3.0	2.06 / 2.84 / 3.07	38.6 / 93.8 / 95.4	50.8 / 62.0 / 49.3	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	3.0	2.32 / 2.89 / 3.05	56.0 / 94.0 / 95.3	76.4 / 83.1 / 77.9	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	3.0	2.66 / 2.95 / 3.02	84.7 / 95.0 / 95.3	95.2 / 95.5 / 95.3	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	3.0	1.23 / 2.46 / 2.73	30.7 / 95.8 / 96.8	0.0 / 0.1 / 0.0	0.6 / 1.1 / 14.2
25-35	25-35	5-15	5-15	3.0	1.32 / 2.38 / 3.01	29.3 / 91.6 / 95.5	0.0 / 0.8 / 0.0	0.0 / 0.0 / 2.6
25-35	45-55	5-15	5-15	3.0	1.46 / 2.47 / 3.18	30.7 / 92.1 / 95.8	0.2 / 3.4 / 0.2	0.0 / 0.0 / 1.0
25-35	65-75	5-15	5-15	3.0	1.57 / 2.51 / 3.13	35.0 / 90.3 / 96.4	0.7 / 8.9 / 2.2	0.0 / 0.0 / 0.3

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
25-35	85-95	5-15	5-15	3.0	1.63 / 2.46 / 3.08	36.9 / 87.2 / 94.2	1.8 / 16.8 / 9.5	0.0 / 0.0 / 0.2
25-35	25-35	5-15	25-35	3.0	1.40 / 2.51 / 3.10	1.5 / 90.8 / 96.5	2.1 / 13.4 / 1.1	0.0 / 0.0 / 0.1
25-35	45-55	5-15	25-35	3.0	1.53 / 2.55 / 3.08	2.1 / 89.8 / 96.1	9.7 / 38.9 / 18.6	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	3.0	1.67 / 2.59 / 3.06	5.7 / 86.7 / 94.2	27.9 / 65.9 / 47.5	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	3.0	1.80 / 2.61 / 3.05	9.1 / 86.5 / 94.9	46.6 / 81.4 / 71.1	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	3.0	1.63 / 2.60 / 3.01	0.5 / 87.9 / 96.4	55.2 / 82.3 / 70.9	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	3.0	1.81 / 2.67 / 3.03	3.5 / 87.2 / 94.6	81.4 / 93.0 / 89.3	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	3.0	1.98 / 2.69 / 3.00	8.1 / 84.0 / 94.1	94.8 / 99.6 / 98.8	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	3.0	2.00 / 2.76 / 3.02	6.6 / 88.2 / 94.2	96.9 / 98.8 / 98.5	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	3.0	2.24 / 2.80 / 3.01	23.9 / 88.3 / 93.4	99.4 / 99.7 / 99.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	3.0	2.62 / 2.93 / 3.02	68.6 / 94.2 / 95.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	3.0	1.20 / 2.34 / 2.47	53.2 / 94.2 / 97.9	0.0 / 0.0 / 0.0	4.0 / 2.3 / 18.3
45-55	25-35	5-15	5-15	3.0	1.35 / 2.33 / 3.03	28.0 / 91.3 / 96.1	0.0 / 1.2 / 0.0	0.3 / 0.0 / 3.0
45-55	45-55	5-15	5-15	3.0	1.42 / 2.31 / 3.08	24.9 / 87.4 / 95.8	0.2 / 3.1 / 0.2	0.1 / 0.0 / 1.0
45-55	65-75	5-15	5-15	3.0	1.49 / 2.34 / 3.09	27.0 / 84.1 / 95.4	0.8 / 10.4 / 4.5	0.1 / 0.0 / 0.3
45-55	85-95	5-15	5-15	3.0	1.62 / 2.36 / 3.05	32.4 / 82.0 / 95.0	2.1 / 17.1 / 12.8	0.1 / 0.0 / 0.0
45-55	25-35	5-15	25-35	3.0	1.39 / 2.43 / 3.08	0.8 / 87.0 / 96.2	2.5 / 16.3 / 3.0	0.0 / 0.0 / 0.0
45-55	45-55	5-15	25-35	3.0	1.52 / 2.48 / 3.12	1.5 / 85.6 / 96.6	14.3 / 44.6 / 26.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	3.0	1.61 / 2.43 / 2.99	2.4 / 79.3 / 94.6	32.2 / 67.9 / 56.3	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	3.0	1.76 / 2.49 / 3.05	5.3 / 77.4 / 95.1	51.2 / 83.7 / 78.8	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	3.0	1.61 / 2.56 / 3.03	0.5 / 83.1 / 96.2	63.1 / 85.4 / 77.5	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	3.0	1.78 / 2.61 / 3.04	1.3 / 81.4 / 97.3	87.0 / 96.9 / 94.8	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	3.0	1.95 / 2.63 / 3.02	6.5 / 77.6 / 94.7	97.2 / 99.5 / 99.3	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	3.0	1.98 / 2.72 / 3.02	3.9 / 88.2 / 94.9	99.0 / 99.6 / 99.6	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	3.0	2.20 / 2.75 / 3.01	16.1 / 85.9 / 93.8	99.7 / 99.9 / 99.9	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	3.0	2.62 / 2.93 / 3.03	64.2 / 93.2 / 94.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	3.0	1.11 / 2.33 / 3.00	0.2 / 91.3 / 96.2	0.2 / 0.7 / 0.0	1.6 / 1.6 / 4.3
25-35	25-35	25-35	5-15	3.0	1.20 / 2.39 / 3.07	0.0 / 86.8 / 97.4	1.0 / 9.9 / 2.3	0.0 / 0.0 / 0.0
25-35	45-55	25-35	5-15	3.0	1.25 / 2.46 / 3.09	0.0 / 85.4 / 94.7	9.5 / 35.1 / 27.4	0.0 / 0.0 / 0.0
25-35	65-75	25-35	5-15	3.0	1.30 / 2.48 / 3.03	0.0 / 82.6 / 94.4	22.7 / 60.2 / 62.4	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	3.0	1.36 / 2.51 / 3.01	0.0 / 81.3 / 96.3	41.5 / 76.1 / 83.2	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	3.0	1.33 / 2.44 / 3.02	0.0 / 77.7 / 96.3	29.1 / 73.2 / 63.6	0.0 / 0.0 / 0.3
25-35	45-55	25-35	25-35	3.0	1.43 / 2.51 / 3.04	0.0 / 74.4 / 95.6	80.5 / 97.9 / 97.4	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	3.0	1.54 / 2.55 / 3.00	0.0 / 73.9 / 94.8	97.4 / 100.0 / 99.9	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	3.0	1.66 / 2.62 / 3.00	0.0 / 74.3 / 94.8	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	3.0	1.58 / 2.57 / 3.01	0.0 / 71.9 / 95.4	99.6 / 100.0 / 99.7	0.0 / 0.0 / 0.1
25-35	65-75	25-35	45-55	3.0	1.74 / 2.65 / 3.01	0.0 / 75.9 / 93.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	3.0	1.92 / 2.73 / 3.01	0.0 / 78.5 / 95.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 1 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	OR	Expected OR	Coverage (%)	Power(%)	Failure rate (%)
25-35	65-75	25-35	65-75	3.0	1.95 / 2.72 / 2.99	0.0 / 77.6 / 94.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.1
25-35	85-95	25-35	65-75	3.0	2.23 / 2.82 / 3.00	0.8 / 86.1 / 95.8	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	3.0	2.59 / 2.92 / 3.00	37.7 / 92.5 / 95.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	3.0	1.13 / 2.25 / 2.86	0.1 / 90.0 / 96.6	0.0 / 0.4 / 0.0	0.0 / 0.0 / 9.4
45-55	25-35	25-35	5-15	3.0	1.19 / 2.32 / 3.14	0.0 / 83.9 / 96.6	1.6 / 11.6 / 1.7	0.0 / 0.0 / 0.0
45-55	45-55	25-35	5-15	3.0	1.23 / 2.30 / 3.02	0.0 / 75.3 / 95.2	8.8 / 36.5 / 30.4	0.0 / 0.0 / 0.0
45-55	65-75	25-35	5-15	3.0	1.27 / 2.34 / 3.02	0.0 / 70.4 / 94.6	23.1 / 60.9 / 68.2	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	3.0	1.33 / 2.39 / 3.04	0.0 / 68.9 / 95.8	43.8 / 81.0 / 89.7	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	3.0	1.32 / 2.39 / 3.04	0.0 / 73.8 / 95.1	32.6 / 79.7 / 76.4	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	3.0	1.41 / 2.41 / 3.01	0.0 / 59.4 / 96.1	81.4 / 99.4 / 99.3	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	3.0	1.50 / 2.45 / 3.00	0.0 / 56.6 / 96.7	98.7 / 99.9 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	3.0	1.62 / 2.52 / 3.01	0.0 / 56.8 / 94.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	3.0	1.55 / 2.51 / 3.01	0.0 / 60.9 / 95.8	99.8 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	3.0	1.71 / 2.58 / 3.00	0.0 / 61.4 / 95.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	3.0	1.87 / 2.64 / 2.99	0.0 / 65.5 / 95.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	3.0	1.94 / 2.70 / 2.99	0.0 / 71.5 / 95.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	3.0	2.18 / 2.76 / 2.99	0.3 / 76.6 / 94.9	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	3.0	2.59 / 2.91 / 3.01	31.6 / 90.0 / 95.4	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	3.0	1.14 / 2.07 / 2.83	3.7 / 88.3 / 96.6	0.0 / 0.3 / 0.0	3.1 / 3.1 / 15.8
45-55	25-35	45-55	5-15	3.0	1.21 / 2.17 / 3.11	0.0 / 76.8 / 95.4	0.6 / 8.5 / 0.6	0.0 / 0.0 / 0.2
45-55	45-55	45-55	5-15	3.0	1.27 / 2.26 / 3.10	0.0 / 73.2 / 96.9	3.1 / 30.1 / 16.6	0.0 / 0.0 / 0.0
45-55	65-75	45-55	5-15	3.0	1.33 / 2.32 / 3.05	0.0 / 70.6 / 95.2	10.5 / 52.1 / 55.1	0.0 / 0.0 / 0.0
45-55	85-95	45-55	5-15	3.0	1.41 / 2.38 / 3.02	0.0 / 69.6 / 93.6	21.1 / 72.6 / 86.1	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	3.0	1.31 / 2.27 / 3.05	0.0 / 55.9 / 94.6	31.8 / 80.6 / 74.1	0.0 / 0.0 / 0.4
45-55	45-55	45-55	25-35	3.0	1.40 / 2.33 / 3.04	0.0 / 47.4 / 94.3	83.0 / 98.8 / 99.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	3.0	1.51 / 2.41 / 3.02	0.0 / 45.5 / 96.0	98.8 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	3.0	1.63 / 2.50 / 3.01	0.0 / 52.2 / 96.6	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	3.0	1.52 / 2.41 / 3.01	0.0 / 39.5 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.3
45-55	65-75	45-55	45-55	3.0	1.68 / 2.51 / 3.02	0.0 / 44.6 / 95.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	3.0	1.86 / 2.60 / 3.01	0.0 / 54.8 / 95.3	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	3.0	1.90 / 2.63 / 3.02	0.0 / 55.9 / 94.7	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	3.0	2.17 / 2.75 / 3.01	0.0 / 71.7 / 96.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	3.0	2.55 / 2.87 / 3.00	21.8 / 88.1 / 94.9	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0

Supplementary Table 2: Cohort simulation study results. NAIVE / DOSAGE / LC

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
5-15	5-15	5-15	5-15	1.0	2.13 / 1.01 / 1.33	88.7 / 99.3 / 93.4	0.0 / 0.0 / 0.0	71.6 / 17.9 / 49.5
5-15	25-35	5-15	5-15	1.0	1.39 / 0.94 / 0.99	93.9 / 97.9 / 95.7	0.0 / 0.0 / 0.0	38.8 / 3.4 / 37.4
5-15	45-55	5-15	5-15	1.0	1.17 / 0.93 / 0.86	94.5 / 94.7 / 97.4	0.0 / 0.0 / 0.0	24.1 / 1.2 / 29.5
5-15	65-75	5-15	5-15	1.0	1.07 / 0.99 / 0.91	96.3 / 96.3 / 97.9	0.0 / 0.0 / 0.0	20.7 / 0.5 / 23.1
5-15	85-95	5-15	5-15	1.0	1.10 / 0.96 / 0.90	94.5 / 95.0 / 97.0	0.0 / 0.0 / 0.0	15.7 / 0.1 / 19.1
5-15	25-35	5-15	25-35	1.0	0.95 / 0.90 / 0.86	95.3 / 95.0 / 97.6	0.0 / 0.0 / 0.0	6.3 / 0.4 / 20.1
5-15	45-55	5-15	25-35	1.0	0.97 / 0.95 / 0.89	95.8 / 95.0 / 97.1	0.0 / 0.0 / 0.0	1.1 / 0.0 / 9.3
5-15	65-75	5-15	25-35	1.0	0.95 / 0.97 / 0.93	95.6 / 94.3 / 96.2	0.0 / 0.0 / 0.0	0.3 / 0.1 / 5.1
5-15	85-95	5-15	25-35	1.0	0.96 / 0.97 / 0.93	96.0 / 95.7 / 96.7	0.0 / 0.0 / 0.0	0.1 / 0.0 / 1.2
5-15	45-55	5-15	45-55	1.0	0.96 / 0.97 / 0.94	95.5 / 95.3 / 96.4	0.0 / 0.0 / 0.0	0.1 / 0.0 / 1.6
5-15	65-75	5-15	45-55	1.0	0.97 / 0.99 / 0.96	95.7 / 95.3 / 96.8	0.0 / 0.0 / 0.0	0.1 / 0.1 / 0.7
5-15	85-95	5-15	45-55	1.0	0.99 / 0.99 / 0.98	95.4 / 95.9 / 96.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.3
5-15	65-75	5-15	65-75	1.0	0.98 / 0.98 / 0.96	96.7 / 95.3 / 96.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
5-15	85-95	5-15	65-75	1.0	0.98 / 0.97 / 0.97	95.1 / 95.2 / 96.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
5-15	85-95	5-15	85-95	1.0	0.98 / 0.98 / 0.98	94.3 / 93.9 / 94.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	1.0	1.02 / 0.95 / 1.07	96.8 / 97.6 / 95.7	0.0 / 0.0 / 0.0	7.1 / 3.6 / 28.5
25-35	25-35	5-15	5-15	1.0	1.02 / 1.00 / 0.94	95.9 / 95.9 / 98.6	0.0 / 0.0 / 0.0	4.7 / 0.0 / 18.2
25-35	45-55	5-15	5-15	1.0	0.97 / 0.98 / 0.94	96.5 / 94.4 / 97.7	0.0 / 0.0 / 0.0	2.6 / 0.1 / 13.2
25-35	65-75	5-15	5-15	1.0	0.99 / 1.00 / 0.95	95.4 / 95.9 / 97.4	0.0 / 0.0 / 0.0	2.3 / 0.0 / 7.0
25-35	85-95	5-15	5-15	1.0	1.03 / 0.99 / 0.96	96.6 / 93.9 / 96.4	0.0 / 0.0 / 0.0	2.9 / 0.0 / 4.6
25-35	25-35	5-15	25-35	1.0	0.99 / 0.98 / 0.94	94.6 / 95.7 / 97.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 3.4
25-35	45-55	5-15	25-35	1.0	0.99 / 1.00 / 0.98	95.1 / 94.6 / 96.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
25-35	65-75	5-15	25-35	1.0	0.98 / 0.98 / 0.97	94.4 / 95.3 / 96.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.3
25-35	85-95	5-15	25-35	1.0	0.98 / 0.99 / 0.98	95.1 / 94.5 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.2
25-35	45-55	5-15	45-55	1.0	0.99 / 0.99 / 0.98	94.8 / 95.7 / 96.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	1.0	0.98 / 0.99 / 0.98	94.9 / 96.0 / 96.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	1.0	1.00 / 1.01 / 1.00	95.9 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	1.0	1.00 / 1.00 / 1.00	95.2 / 95.4 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	1.0	0.99 / 0.99 / 0.99	94.9 / 94.4 / 94.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	1.0	1.00 / 1.00 / 1.00	96.7 / 96.6 / 96.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	1.0	1.04 / 0.96 / 0.94	92.6 / 97.1 / 99.3	0.0 / 0.0 / 0.0	7.6 / 3.8 / 26.0
45-55	25-35	5-15	5-15	1.0	1.03 / 1.02 / 1.04	95.6 / 95.7 / 99.5	0.0 / 0.0 / 0.0	1.1 / 0.2 / 15.5
45-55	45-55	5-15	5-15	1.0	1.01 / 1.04 / 1.02	95.4 / 95.5 / 97.5	0.0 / 0.0 / 0.0	1.0 / 0.0 / 8.0
45-55	65-75	5-15	5-15	1.0	0.98 / 0.98 / 0.98	96.1 / 96.1 / 97.5	0.0 / 0.0 / 0.0	1.6 / 0.0 / 5.2
45-55	85-95	5-15	5-15	1.0	1.00 / 1.03 / 1.01	96.2 / 94.4 / 96.1	0.0 / 0.0 / 0.0	1.4 / 0.0 / 4.5
45-55	25-35	5-15	25-35	1.0	1.01 / 1.03 / 1.03	95.1 / 95.7 / 97.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 1.9
45-55	45-55	5-15	25-35	1.0	1.00 / 1.00 / 1.01	94.2 / 95.0 / 96.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	5-15	25-35	1.0	1.00 / 1.00 / 1.00	95.6 / 95.2 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
45-55	85-95	5-15	25-35	1.0	1.01 / 1.02 / 1.01	95.1 / 95.6 / 95.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	1.0	1.01 / 1.01 / 1.01	95.3 / 94.9 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	1.0	1.00 / 1.00 / 1.00	95.9 / 95.2 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	1.0	1.00 / 1.00 / 1.00	94.8 / 95.3 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	1.0	1.00 / 1.00 / 1.00	94.8 / 95.3 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	1.0	0.99 / 0.99 / 0.99	95.2 / 93.8 / 94.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	1.0	1.00 / 1.00 / 1.00	95.0 / 95.7 / 95.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	1.0	0.99 / 0.96 / 0.96	95.2 / 94.5 / 95.4	0.0 / 0.0 / 0.1	0.1 / 0.1 / 13.9
25-35	25-35	25-35	5-15	1.0	1.00 / 1.02 / 0.98	94.4 / 94.7 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 4.2
25-35	45-55	25-35	5-15	1.0	1.00 / 1.00 / 0.97	95.7 / 94.8 / 93.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 2.3
25-35	65-75	25-35	5-15	1.0	1.00 / 0.99 / 0.98	95.1 / 94.9 / 92.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 1.8
25-35	85-95	25-35	5-15	1.0	1.00 / 1.00 / 0.99	96.3 / 96.1 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.6
25-35	25-35	25-35	25-35	1.0	0.99 / 0.99 / 0.98	95.2 / 95.5 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	25-35	1.0	1.00 / 1.01 / 1.00	94.6 / 95.3 / 95.5	0.0 / 0.1 / 0.1	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.6 / 95.9 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	1.0	1.00 / 1.00 / 1.00	96.1 / 95.1 / 94.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.0 / 94.5 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
25-35	65-75	25-35	45-55	1.0	1.00 / 1.00 / 1.00	95.1 / 94.3 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	1.0	0.99 / 0.99 / 0.99	95.6 / 95.2 / 95.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	1.0	1.00 / 1.00 / 1.00	95.2 / 95.0 / 95.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	1.0	1.00 / 1.00 / 1.00	95.1 / 94.5 / 94.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	1.0	1.00 / 1.00 / 1.00	95.0 / 94.9 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	1.0	1.00 / 1.00 / 1.01	93.6 / 95.4 / 95.4	0.0 / 0.0 / 0.0	0.0 / 0.4 / 13.5
45-55	25-35	25-35	5-15	1.0	0.99 / 0.99 / 0.99	95.6 / 94.4 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 4.8
45-55	45-55	25-35	5-15	1.0	1.01 / 1.02 / 1.01	95.3 / 96.2 / 94.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 1.9
45-55	65-75	25-35	5-15	1.0	1.00 / 1.00 / 0.99	94.9 / 94.9 / 93.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.9
45-55	85-95	25-35	5-15	1.0	0.99 / 1.00 / 1.00	94.6 / 94.7 / 92.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.5
45-55	25-35	25-35	25-35	1.0	1.00 / 1.00 / 1.00	95.1 / 94.9 / 93.8	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	1.0	1.00 / 0.99 / 0.99	94.3 / 95.0 / 94.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	1.0	1.00 / 1.00 / 1.00	94.9 / 94.4 / 94.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	1.0	1.00 / 1.00 / 1.00	95.3 / 94.4 / 94.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	1.0	1.00 / 1.00 / 1.00	96.2 / 95.5 / 95.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	1.0	1.00 / 1.00 / 1.00	94.5 / 94.6 / 94.4	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	1.0	1.00 / 1.00 / 1.00	94.7 / 94.0 / 94.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	1.0	1.01 / 1.00 / 1.00	94.0 / 94.6 / 94.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	1.0	1.00 / 1.00 / 1.00	94.1 / 94.0 / 94.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	1.0	1.00 / 1.00 / 1.00	94.1 / 94.0 / 94.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	5-15	45-55	5-15	1.0	1.01 / 1.07 / 1.04	94.9 / 95.3 / 95.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 21.2
45-55	25-35	45-55	5-15	1.0	1.00 / 1.00 / 1.01	95.3 / 96.0 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 3.9
45-55	45-55	45-55	5-15	1.0	1.01 / 0.99 / 1.00	95.4 / 95.8 / 94.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 2.2
45-55	65-75	45-55	5-15	1.0	1.01 / 1.00 / 1.00	94.1 / 94.7 / 91.9	0.0 / 0.0 / 0.1	0.0 / 0.0 / 1.6
45-55	85-95	45-55	5-15	1.0	1.01 / 1.00 / 1.00	96.0 / 95.7 / 91.9	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.9
45-55	25-35	45-55	25-35	1.0	1.00 / 1.00 / 1.00	94.8 / 95.6 / 95.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.1
45-55	45-55	45-55	25-35	1.0	1.00 / 1.01 / 1.01	95.0 / 94.9 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	1.0	1.00 / 1.00 / 1.00	95.5 / 94.9 / 94.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	1.0	1.00 / 1.00 / 1.00	94.4 / 94.5 / 94.1	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	1.0	1.00 / 1.00 / 1.00	94.8 / 94.2 / 94.3	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	45-55	1.0	1.00 / 1.00 / 1.00	95.8 / 95.2 / 95.2	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	1.0	1.00 / 1.00 / 1.00	93.9 / 94.5 / 94.6	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	1.0	1.00 / 1.00 / 1.00	95.1 / 94.8 / 94.7	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	1.0	1.00 / 1.00 / 1.00	95.5 / 95.1 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	1.0	1.00 / 1.00 / 1.00	95.6 / 94.9 / 95.0	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	1.5	2.14 / 1.20 / 1.61	94.7 / 99.9 / 95.7	0.0 / 0.0 / 0.8	67.5 / 18.9 / 42.2
5-15	25-35	5-15	5-15	1.5	1.42 / 1.44 / 1.37	97.9 / 98.1 / 97.5	0.0 / 0.0 / 0.1	31.4 / 4.1 / 27.0
5-15	45-55	5-15	5-15	1.5	1.33 / 1.55 / 1.42	97.5 / 96.1 / 97.9	0.1 / 0.0 / 0.0	18.7 / 0.4 / 14.3
5-15	65-75	5-15	5-15	1.5	1.33 / 1.54 / 1.42	97.1 / 94.8 / 96.7	0.0 / 0.0 / 0.0	15.5 / 0.8 / 10.9
5-15	85-95	5-15	5-15	1.5	1.30 / 1.52 / 1.40	96.9 / 94.8 / 97.7	0.1 / 0.0 / 0.1	12.6 / 0.2 / 6.9
5-15	25-35	5-15	25-35	1.5	1.12 / 1.46 / 1.32	96.1 / 94.7 / 97.4	0.0 / 0.1 / 0.0	3.5 / 0.1 / 7.6
5-15	45-55	5-15	25-35	1.5	1.21 / 1.53 / 1.42	94.0 / 95.3 / 96.8	0.0 / 0.1 / 0.0	0.3 / 0.0 / 2.1
5-15	65-75	5-15	25-35	1.5	1.22 / 1.49 / 1.42	94.0 / 96.1 / 97.4	0.0 / 0.1 / 0.2	0.4 / 0.0 / 1.1
5-15	85-95	5-15	25-35	1.5	1.26 / 1.50 / 1.44	92.7 / 94.4 / 96.7	0.1 / 0.1 / 0.1	0.1 / 0.0 / 0.4
5-15	45-55	5-15	45-55	1.5	1.23 / 1.49 / 1.44	90.8 / 94.5 / 95.7	0.3 / 0.3 / 0.1	0.0 / 0.0 / 0.2
5-15	65-75	5-15	45-55	1.5	1.28 / 1.49 / 1.46	92.8 / 95.1 / 96.6	0.2 / 0.2 / 0.2	0.0 / 0.0 / 0.3
5-15	85-95	5-15	45-55	1.5	1.32 / 1.49 / 1.46	92.9 / 94.7 / 95.5	0.3 / 0.5 / 0.2	0.0 / 0.0 / 0.2
5-15	65-75	5-15	65-75	1.5	1.32 / 1.49 / 1.46	92.2 / 94.5 / 95.4	0.2 / 0.5 / 0.4	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	1.5	1.38 / 1.49 / 1.47	94.4 / 95.6 / 96.0	0.7 / 0.7 / 1.1	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	1.5	1.44 / 1.49 / 1.49	95.7 / 95.7 / 95.7	1.4 / 1.5 / 1.5	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	1.5	1.10 / 1.56 / 1.47	91.5 / 98.6 / 97.4	0.0 / 0.0 / 0.3	5.3 / 2.9 / 19.6
25-35	25-35	5-15	5-15	1.5	1.13 / 1.56 / 1.44	89.8 / 95.2 / 97.8	0.0 / 0.0 / 0.0	3.4 / 0.2 / 8.2
25-35	45-55	5-15	5-15	1.5	1.18 / 1.60 / 1.46	92.9 / 94.0 / 96.9	0.0 / 0.0 / 0.0	2.2 / 0.1 / 5.5
25-35	65-75	5-15	5-15	1.5	1.21 / 1.60 / 1.48	90.3 / 95.2 / 96.5	0.0 / 0.0 / 0.0	1.4 / 0.0 / 2.0
25-35	85-95	5-15	5-15	1.5	1.28 / 1.61 / 1.47	93.6 / 95.3 / 96.1	0.0 / 0.2 / 0.1	1.9 / 0.0 / 0.6
25-35	25-35	5-15	25-35	1.5	1.19 / 1.61 / 1.54	83.3 / 95.1 / 97.5	0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.5
25-35	45-55	5-15	25-35	1.5	1.21 / 1.56 / 1.49	84.5 / 93.3 / 95.4	0.1 / 0.3 / 0.2	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	1.5	1.26 / 1.54 / 1.48	85.0 / 94.3 / 94.7	0.2 / 0.4 / 0.3	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
25-35	85-95	5-15	25-35	1.5	1.29 / 1.55 / 1.50	89.6 / 94.4 / 95.2	0.3 / 0.8 / 0.3	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	1.5	1.25 / 1.55 / 1.50	79.6 / 95.9 / 96.6	0.2 / 0.6 / 0.3	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	1.5	1.30 / 1.54 / 1.50	82.9 / 96.2 / 96.4	0.4 / 1.2 / 1.1	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	1.5	1.34 / 1.52 / 1.49	87.3 / 93.2 / 93.2	1.4 / 3.0 / 3.0	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	1.5	1.34 / 1.53 / 1.51	86.0 / 95.1 / 95.5	1.7 / 2.8 / 2.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	1.5	1.40 / 1.52 / 1.50	90.0 / 94.4 / 94.9	5.2 / 6.6 / 7.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	1.5	1.44 / 1.50 / 1.49	93.9 / 95.3 / 95.2	12.9 / 14.0 / 14.1	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	1.5	1.13 / 1.46 / 1.48	89.7 / 96.7 / 97.6	0.0 / 0.0 / 0.0	7.8 / 4.2 / 19.7
45-55	25-35	5-15	5-15	1.5	1.15 / 1.63 / 1.55	90.3 / 95.7 / 98.1	0.0 / 0.0 / 0.0	0.9 / 0.1 / 5.3
45-55	45-55	5-15	5-15	1.5	1.16 / 1.60 / 1.49	90.9 / 95.2 / 96.5	0.0 / 0.0 / 0.0	0.9 / 0.0 / 2.8
45-55	65-75	5-15	5-15	1.5	1.25 / 1.64 / 1.54	91.8 / 94.8 / 95.2	0.0 / 0.1 / 0.0	1.7 / 0.0 / 2.3
45-55	85-95	5-15	5-15	1.5	1.26 / 1.60 / 1.49	91.6 / 95.7 / 95.5	0.0 / 0.2 / 0.0	1.3 / 0.0 / 0.7
45-55	25-35	5-15	25-35	1.5	1.19 / 1.57 / 1.52	79.4 / 94.4 / 95.8	0.0 / 0.1 / 0.0	0.0 / 0.0 / 0.4
45-55	45-55	5-15	25-35	1.5	1.22 / 1.56 / 1.51	77.0 / 94.2 / 95.2	0.2 / 0.3 / 0.2	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	1.5	1.26 / 1.56 / 1.50	82.5 / 93.9 / 95.1	0.2 / 0.8 / 0.2	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	1.5	1.30 / 1.56 / 1.50	86.7 / 94.4 / 95.1	0.0 / 0.9 / 0.6	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	1.5	1.25 / 1.53 / 1.51	73.4 / 94.6 / 95.3	0.2 / 0.7 / 0.2	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	1.5	1.30 / 1.53 / 1.50	78.2 / 94.8 / 95.7	0.5 / 1.1 / 0.7	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	1.5	1.35 / 1.54 / 1.50	86.0 / 94.7 / 94.3	1.8 / 3.0 / 2.3	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	1.5	1.34 / 1.51 / 1.50	84.3 / 95.1 / 95.5	1.9 / 3.1 / 2.7	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	1.5	1.41 / 1.54 / 1.52	91.7 / 95.3 / 95.5	6.3 / 8.5 / 7.9	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	1.5	1.45 / 1.51 / 1.50	94.4 / 95.1 / 94.7	13.5 / 16.4 / 15.8	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	1.5	1.06 / 1.59 / 1.48	44.1 / 94.7 / 94.6	0.0 / 0.0 / 1.1	0.0 / 0.0 / 11.8
25-35	25-35	25-35	5-15	1.5	1.10 / 1.55 / 1.47	40.5 / 93.9 / 93.0	0.0 / 0.0 / 0.2	0.0 / 0.0 / 1.5
25-35	45-55	25-35	5-15	1.5	1.11 / 1.52 / 1.46	33.0 / 94.8 / 94.8	0.0 / 0.1 / 0.2	0.0 / 0.0 / 0.6
25-35	65-75	25-35	5-15	1.5	1.13 / 1.51 / 1.46	34.4 / 94.6 / 94.9	0.3 / 0.3 / 1.0	0.0 / 0.0 / 0.1
25-35	85-95	25-35	5-15	1.5	1.16 / 1.54 / 1.49	38.6 / 93.9 / 94.2	0.1 / 0.5 / 1.5	0.0 / 0.0 / 0.1
25-35	25-35	25-35	25-35	1.5	1.15 / 1.56 / 1.52	39.7 / 94.5 / 94.1	0.1 / 0.5 / 0.5	0.0 / 0.0 / 0.0
25-35	45-55	25-35	25-35	1.5	1.19 / 1.53 / 1.50	43.6 / 96.1 / 95.8	0.1 / 0.4 / 0.9	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	1.5	1.22 / 1.52 / 1.49	49.3 / 94.3 / 94.8	0.8 / 2.3 / 2.4	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	1.5	1.26 / 1.52 / 1.50	58.3 / 93.4 / 94.6	2.6 / 5.8 / 6.8	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	1.5	1.23 / 1.52 / 1.50	46.0 / 94.7 / 95.3	2.0 / 4.2 / 4.2	0.0 / 0.0 / 0.0
25-35	65-75	25-35	45-55	1.5	1.27 / 1.50 / 1.49	56.8 / 94.0 / 93.8	4.0 / 8.5 / 8.6	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	1.5	1.32 / 1.51 / 1.50	68.4 / 94.6 / 94.5	11.7 / 17.2 / 17.7	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	1.5	1.33 / 1.51 / 1.50	71.9 / 95.5 / 95.5	13.7 / 19.6 / 19.8	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	1.5	1.38 / 1.50 / 1.49	83.4 / 94.3 / 94.8	28.9 / 35.0 / 35.2	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	1.5	1.44 / 1.50 / 1.50	93.5 / 96.0 / 96.1	53.6 / 55.8 / 56.5	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	1.5	1.08 / 1.50 / 1.42	69.3 / 95.1 / 93.8	0.0 / 0.0 / 0.2	0.0 / 0.2 / 15.4

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	25-35	25-35	5-15	1.5	1.09 / 1.52 / 1.49	34.9 / 95.7 / 94.7	0.0 / 0.1 / 0.1	0.0 / 0.0 / 2.6
45-55	45-55	25-35	5-15	1.5	1.11 / 1.51 / 1.48	27.0 / 95.7 / 94.9	0.0 / 0.1 / 0.4	0.0 / 0.0 / 0.7
45-55	65-75	25-35	5-15	1.5	1.14 / 1.55 / 1.50	31.4 / 94.6 / 94.0	0.0 / 0.6 / 1.3	0.0 / 0.0 / 0.3
45-55	85-95	25-35	5-15	1.5	1.16 / 1.54 / 1.50	35.3 / 95.6 / 95.0	0.1 / 0.2 / 1.8	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	1.5	1.15 / 1.52 / 1.51	36.7 / 95.3 / 94.6	0.1 / 0.4 / 0.8	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	1.5	1.18 / 1.49 / 1.48	32.4 / 94.6 / 94.7	0.0 / 1.0 / 1.3	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	1.5	1.22 / 1.52 / 1.51	42.9 / 95.0 / 95.1	1.0 / 2.4 / 3.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	1.5	1.26 / 1.52 / 1.51	52.7 / 94.9 / 94.8	3.2 / 5.8 / 7.1	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	1.5	1.22 / 1.50 / 1.50	38.2 / 95.8 / 95.2	2.1 / 3.2 / 4.2	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	1.5	1.28 / 1.51 / 1.51	53.3 / 94.3 / 95.3	6.0 / 10.8 / 11.9	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	1.5	1.32 / 1.51 / 1.50	67.5 / 95.6 / 95.5	11.4 / 19.0 / 20.4	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	1.5	1.32 / 1.50 / 1.50	67.8 / 95.7 / 96.0	16.4 / 23.2 / 23.9	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	1.5	1.38 / 1.50 / 1.50	82.2 / 94.6 / 94.2	34.2 / 40.6 / 40.8	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	1.5	1.44 / 1.50 / 1.50	92.7 / 95.1 / 95.7	57.9 / 60.7 / 61.5	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	1.5	1.07 / 1.53 / 1.46	80.5 / 95.9 / 94.3	0.0 / 0.0 / 0.2	0.0 / 0.2 / 17.0
45-55	25-35	45-55	5-15	1.5	1.10 / 1.45 / 1.44	66.9 / 93.4 / 91.8	0.0 / 0.2 / 0.4	0.0 / 0.0 / 3.7
45-55	45-55	45-55	5-15	1.5	1.10 / 1.44 / 1.42	58.2 / 94.5 / 93.5	0.1 / 0.1 / 0.4	0.0 / 0.0 / 1.5
45-55	65-75	45-55	5-15	1.5	1.16 / 1.48 / 1.47	66.3 / 94.5 / 94.2	0.1 / 0.1 / 1.2	0.0 / 0.0 / 1.1
45-55	85-95	45-55	5-15	1.5	1.17 / 1.47 / 1.47	68.4 / 94.9 / 94.0	0.0 / 0.2 / 1.3	0.0 / 0.0 / 0.5
45-55	25-35	45-55	25-35	1.5	1.14 / 1.47 / 1.50	29.1 / 95.7 / 95.4	0.2 / 0.4 / 0.6	0.0 / 0.0 / 0.0
45-55	45-55	45-55	25-35	1.5	1.17 / 1.46 / 1.49	28.0 / 95.6 / 96.0	0.1 / 0.5 / 0.8	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	1.5	1.20 / 1.47 / 1.49	35.3 / 94.1 / 94.7	0.9 / 1.5 / 3.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	1.5	1.24 / 1.48 / 1.50	45.7 / 96.2 / 95.7	1.9 / 3.7 / 6.2	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	1.5	1.20 / 1.46 / 1.48	26.5 / 94.2 / 94.7	1.0 / 2.7 / 3.6	0.0 / 0.0 / 0.0
45-55	65-75	45-55	45-55	1.5	1.26 / 1.48 / 1.50	43.7 / 95.4 / 95.5	5.0 / 9.4 / 11.3	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	1.5	1.30 / 1.48 / 1.49	57.1 / 95.0 / 95.0	11.3 / 17.4 / 19.2	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	1.5	1.31 / 1.49 / 1.51	61.7 / 95.0 / 95.5	16.3 / 23.8 / 26.2	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	1.5	1.37 / 1.49 / 1.50	78.6 / 95.5 / 95.8	33.4 / 42.3 / 44.1	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	1.5	1.43 / 1.49 / 1.50	91.9 / 95.3 / 95.5	57.9 / 61.0 / 61.8	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	2.0	2.12 / 1.63 / 1.79	95.3 / 99.7 / 95.4	0.0 / 0.0 / 0.9	61.4 / 17.2 / 34.2
5-15	25-35	5-15	5-15	2.0	1.56 / 2.06 / 1.83	97.3 / 97.1 / 97.5	0.0 / 0.0 / 0.6	26.3 / 5.6 / 17.9
5-15	45-55	5-15	5-15	2.0	1.49 / 2.20 / 1.93	92.7 / 95.1 / 97.4	0.0 / 0.0 / 0.2	15.7 / 1.2 / 9.3
5-15	65-75	5-15	5-15	2.0	1.41 / 2.12 / 1.81	92.2 / 93.7 / 97.2	0.1 / 0.2 / 0.2	12.1 / 0.3 / 6.6
5-15	85-95	5-15	5-15	2.0	1.53 / 2.23 / 1.96	94.4 / 93.4 / 96.9	0.0 / 0.5 / 0.8	8.6 / 0.1 / 2.9
5-15	25-35	5-15	25-35	2.0	1.34 / 2.20 / 1.91	90.8 / 94.3 / 97.8	0.3 / 0.2 / 0.3	1.4 / 0.2 / 2.8
5-15	45-55	5-15	25-35	2.0	1.39 / 2.19 / 1.96	83.2 / 94.6 / 97.4	0.1 / 0.2 / 0.2	0.4 / 0.0 / 1.0
5-15	65-75	5-15	25-35	2.0	1.44 / 2.11 / 1.94	82.8 / 93.4 / 97.1	0.9 / 1.4 / 0.8	0.1 / 0.0 / 0.3
5-15	85-95	5-15	25-35	2.0	1.56 / 2.13 / 1.98	86.9 / 94.4 / 96.1	1.2 / 3.1 / 2.7	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
5-15	45-55	5-15	45-55	2.0	1.48 / 2.13 / 1.97	75.3 / 95.9 / 97.5	0.6 / 2.1 / 2.0	0.0 / 0.0 / 0.1
5-15	65-75	5-15	45-55	2.0	1.58 / 2.12 / 2.00	79.5 / 93.4 / 95.1	3.1 / 6.8 / 5.2	0.0 / 0.0 / 0.1
5-15	85-95	5-15	45-55	2.0	1.68 / 2.09 / 2.01	87.4 / 93.8 / 95.4	8.2 / 13.6 / 11.5	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	2.0	1.66 / 2.06 / 1.98	84.4 / 95.4 / 96.2	7.5 / 10.8 / 9.7	0.0 / 0.0 / 0.0
5-15	85-95	5-15	65-75	2.0	1.78 / 2.05 / 2.00	91.0 / 94.6 / 94.7	17.9 / 22.7 / 21.5	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	2.0	1.90 / 2.02 / 2.01	94.6 / 93.6 / 94.4	35.7 / 38.8 / 38.6	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	2.0	1.17 / 2.12 / 1.95	78.0 / 97.9 / 97.0	0.0 / 0.0 / 0.4	4.1 / 4.8 / 20.7
25-35	25-35	5-15	5-15	2.0	1.27 / 2.32 / 1.97	78.6 / 95.1 / 96.7	0.0 / 0.0 / 0.3	2.2 / 0.5 / 4.0
25-35	45-55	5-15	5-15	2.0	1.33 / 2.28 / 1.99	79.9 / 93.3 / 95.9	0.0 / 0.4 / 0.4	1.2 / 0.2 / 1.5
25-35	65-75	5-15	5-15	2.0	1.44 / 2.37 / 2.01	80.4 / 92.8 / 95.0	0.1 / 1.4 / 1.7	1.5 / 0.0 / 0.2
25-35	85-95	5-15	5-15	2.0	1.49 / 2.38 / 2.04	84.1 / 92.9 / 94.5	0.0 / 2.1 / 3.4	1.2 / 0.0 / 0.7
25-35	25-35	5-15	25-35	2.0	1.32 / 2.20 / 2.02	50.5 / 94.3 / 95.3	0.2 / 1.2 / 1.5	0.0 / 0.0 / 0.2
25-35	45-55	5-15	25-35	2.0	1.41 / 2.18 / 2.01	54.2 / 95.6 / 95.8	0.5 / 3.6 / 2.2	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	2.0	1.49 / 2.17 / 2.01	63.8 / 93.7 / 94.1	2.0 / 9.0 / 7.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	2.0	1.58 / 2.16 / 2.00	72.6 / 93.5 / 95.6	5.2 / 16.8 / 14.2	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	2.0	1.47 / 2.09 / 2.00	41.8 / 94.1 / 94.7	5.9 / 15.5 / 12.7	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	2.0	1.58 / 2.09 / 2.01	56.6 / 92.8 / 93.8	18.4 / 33.5 / 29.2	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	2.0	1.68 / 2.10 / 2.01	71.7 / 94.4 / 95.2	32.9 / 52.2 / 49.8	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	2.0	1.66 / 2.07 / 2.02	67.5 / 95.3 / 95.3	41.6 / 52.9 / 50.8	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	2.0	1.77 / 2.03 / 2.00	80.4 / 94.8 / 95.2	57.1 / 66.4 / 65.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	2.0	1.87 / 2.00 / 1.99	90.1 / 94.2 / 94.6	84.1 / 85.9 / 85.4	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	2.0	1.23 / 2.15 / 1.99	83.9 / 97.0 / 96.4	0.0 / 0.0 / 0.0	6.1 / 4.4 / 16.6
45-55	25-35	5-15	5-15	2.0	1.30 / 2.35 / 2.02	72.7 / 94.6 / 96.1	0.0 / 0.0 / 0.0	0.7 / 0.0 / 3.7
45-55	45-55	5-15	5-15	2.0	1.33 / 2.34 / 2.03	72.2 / 93.5 / 95.5	0.0 / 0.5 / 0.0	0.9 / 0.2 / 1.3
45-55	65-75	5-15	5-15	2.0	1.41 / 2.37 / 2.03	76.1 / 92.5 / 93.6	0.0 / 1.0 / 2.0	0.7 / 0.0 / 0.3
45-55	85-95	5-15	5-15	2.0	1.49 / 2.35 / 2.01	81.0 / 93.3 / 95.4	0.1 / 1.8 / 2.0	0.5 / 0.0 / 0.1
45-55	25-35	5-15	25-35	2.0	1.32 / 2.15 / 2.05	43.9 / 95.5 / 95.9	0.3 / 0.9 / 0.6	0.0 / 0.0 / 0.1
45-55	45-55	5-15	25-35	2.0	1.39 / 2.12 / 2.00	45.3 / 95.8 / 95.8	0.5 / 3.9 / 2.1	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	2.0	1.52 / 2.20 / 2.04	59.2 / 92.3 / 94.1	3.2 / 11.7 / 9.5	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	2.0	1.60 / 2.18 / 2.02	69.0 / 94.3 / 95.6	6.7 / 21.9 / 19.6	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	2.0	1.45 / 2.07 / 2.03	30.7 / 93.6 / 94.6	6.7 / 17.7 / 13.4	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	2.0	1.55 / 2.06 / 2.01	45.1 / 95.2 / 95.4	19.0 / 37.0 / 34.6	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	2.0	1.68 / 2.09 / 2.02	69.4 / 94.0 / 95.7	41.0 / 59.3 / 58.6	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	2.0	1.64 / 2.02 / 2.00	55.5 / 94.5 / 95.2	47.2 / 57.9 / 55.7	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	2.0	1.77 / 2.04 / 2.01	80.3 / 95.0 / 94.3	71.4 / 78.1 / 76.9	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	2.0	1.88 / 2.01 / 2.01	90.1 / 95.5 / 96.0	92.9 / 94.0 / 93.9	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	2.0	1.11 / 2.03 / 1.88	4.8 / 94.0 / 91.5	0.0 / 0.1 / 2.5	0.3 / 0.4 / 11.4
25-35	25-35	25-35	5-15	2.0	1.17 / 2.07 / 1.99	2.8 / 94.0 / 92.9	0.1 / 0.5 / 3.7	0.0 / 0.0 / 1.7

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
25-35	45-55	25-35	5-15	2.0	1.20 / 2.04 / 2.00	1.2 / 95.1 / 94.0	0.3 / 1.7 / 6.3	0.0 / 0.0 / 0.3
25-35	65-75	25-35	5-15	2.0	1.24 / 2.02 / 1.99	2.1 / 95.6 / 94.8	1.0 / 3.0 / 10.2	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	2.0	1.28 / 2.03 / 1.99	3.4 / 96.0 / 96.2	2.5 / 7.3 / 17.5	0.0 / 0.0 / 0.1
25-35	25-35	25-35	25-35	2.0	1.26 / 1.98 / 1.99	3.5 / 95.7 / 95.1	1.6 / 3.5 / 6.7	0.0 / 0.0 / 0.0
25-35	45-55	25-35	25-35	2.0	1.33 / 1.98 / 2.00	2.2 / 94.5 / 94.1	5.5 / 17.6 / 22.5	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	2.0	1.40 / 1.99 / 2.00	5.5 / 94.1 / 95.3	17.3 / 40.0 / 42.7	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	2.0	1.47 / 1.99 / 2.01	10.1 / 95.5 / 95.4	35.2 / 62.6 / 67.7	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	2.0	1.41 / 1.97 / 2.00	4.1 / 94.9 / 94.5	25.0 / 51.0 / 52.4	0.0 / 0.0 / 0.0
25-35	65-75	25-35	45-55	2.0	1.50 / 1.99 / 2.02	9.8 / 96.3 / 96.2	60.5 / 81.4 / 82.8	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	2.0	1.58 / 1.98 / 2.00	23.8 / 94.7 / 95.0	82.0 / 93.8 / 94.6	0.0 / 0.0 / 0.0
25-35	65-75	25-35	65-75	2.0	1.61 / 2.00 / 2.01	25.9 / 95.2 / 94.8	91.0 / 96.1 / 96.6	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	2.0	1.73 / 2.00 / 2.01	54.0 / 94.3 / 94.5	99.1 / 99.8 / 99.9	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	2.0	1.87 / 2.00 / 2.01	83.8 / 94.6 / 94.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	2.0	1.11 / 1.87 / 1.87	31.2 / 95.1 / 90.4	0.0 / 0.0 / 1.5	0.0 / 0.4 / 14.1
45-55	25-35	25-35	5-15	2.0	1.15 / 1.91 / 2.00	1.1 / 94.0 / 92.4	0.0 / 0.2 / 3.0	0.0 / 0.0 / 1.9
45-55	45-55	25-35	5-15	2.0	1.20 / 1.98 / 2.01	1.0 / 94.5 / 93.5	0.6 / 1.4 / 6.8	0.0 / 0.0 / 0.1
45-55	65-75	25-35	5-15	2.0	1.24 / 1.98 / 1.99	2.0 / 94.8 / 95.7	0.9 / 2.9 / 12.7	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	2.0	1.28 / 2.00 / 1.99	3.0 / 95.5 / 94.2	2.1 / 5.2 / 22.4	0.0 / 0.0 / 0.1
45-55	25-35	25-35	25-35	2.0	1.25 / 1.91 / 2.03	1.9 / 95.3 / 94.0	1.5 / 4.4 / 7.3	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	2.0	1.30 / 1.91 / 2.00	1.1 / 94.4 / 96.0	5.0 / 15.1 / 21.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	2.0	1.38 / 1.94 / 2.01	2.6 / 92.9 / 93.6	14.2 / 35.8 / 45.1	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	2.0	1.45 / 1.96 / 1.99	9.0 / 94.0 / 94.7	32.4 / 53.2 / 63.8	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	2.0	1.38 / 1.91 / 2.02	2.2 / 92.5 / 95.3	23.2 / 45.1 / 52.2	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	2.0	1.47 / 1.92 / 2.00	5.3 / 93.8 / 95.1	52.5 / 73.2 / 79.9	0.0 / 0.0 / 0.1
45-55	85-95	25-35	45-55	2.0	1.57 / 1.95 / 1.99	18.1 / 94.8 / 95.7	82.3 / 91.2 / 93.5	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	2.0	1.57 / 1.92 / 1.99	15.0 / 93.2 / 94.7	88.3 / 94.9 / 96.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	2.0	1.71 / 1.97 / 2.00	51.5 / 94.3 / 94.0	98.2 / 99.3 / 99.4	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	2.0	1.85 / 1.97 / 2.00	81.4 / 94.2 / 94.7	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	2.0	1.11 / 1.69 / 1.84	58.1 / 95.6 / 94.3	0.0 / 0.0 / 0.7	0.0 / 0.1 / 16.1
45-55	25-35	45-55	5-15	2.0	1.14 / 1.71 / 1.84	24.0 / 93.2 / 94.8	0.0 / 0.1 / 1.0	0.0 / 0.0 / 6.9
45-55	45-55	45-55	5-15	2.0	1.18 / 1.72 / 1.91	18.1 / 93.4 / 94.3	0.0 / 0.2 / 3.7	0.0 / 0.0 / 3.0
45-55	65-75	45-55	5-15	2.0	1.23 / 1.77 / 1.93	24.4 / 93.1 / 94.4	0.2 / 0.9 / 11.0	0.0 / 0.0 / 0.2
45-55	85-95	45-55	5-15	2.0	1.27 / 1.79 / 1.93	29.3 / 92.5 / 94.6	0.3 / 2.7 / 17.9	0.0 / 0.0 / 0.3
45-55	25-35	45-55	25-35	2.0	1.19 / 1.72 / 1.97	0.7 / 91.4 / 94.2	0.5 / 1.4 / 7.5	0.0 / 0.0 / 0.1
45-55	45-55	45-55	25-35	2.0	1.26 / 1.77 / 1.98	0.4 / 89.7 / 95.0	2.1 / 7.1 / 16.4	0.0 / 0.0 / 0.1
45-55	65-75	45-55	25-35	2.0	1.32 / 1.79 / 1.99	0.9 / 89.7 / 95.4	7.4 / 18.5 / 32.8	0.0 / 0.0 / 0.1
45-55	85-95	45-55	25-35	2.0	1.40 / 1.86 / 2.02	4.6 / 92.1 / 94.5	20.5 / 40.0 / 56.4	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	2.0	1.33 / 1.81 / 2.00	0.1 / 90.5 / 94.8	13.1 / 27.3 / 39.7	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	45-55	45-55	2.0	1.42 / 1.85 / 2.00	2.8 / 89.9 / 94.3	39.0 / 58.9 / 69.4	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	2.0	1.51 / 1.88 / 2.00	10.4 / 91.3 / 96.0	64.5 / 79.5 / 85.7	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	2.0	1.54 / 1.88 / 2.00	12.8 / 89.7 / 93.9	76.5 / 86.7 / 90.2	0.0 / 0.0 / 0.1
45-55	85-95	45-55	65-75	2.0	1.67 / 1.92 / 2.01	41.7 / 92.4 / 94.1	94.3 / 97.0 / 97.8	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	2.0	1.83 / 1.96 / 2.01	81.0 / 93.4 / 95.1	99.7 / 99.9 / 99.8	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	2.5	2.18 / 2.10 / 2.40	97.5 / 99.3 / 93.6	0.2 / 0.0 / 1.1	59.1 / 21.1 / 33.2
5-15	25-35	5-15	5-15	2.5	1.67 / 2.95 / 2.42	94.1 / 98.3 / 95.0	0.0 / 0.1 / 2.0	23.7 / 5.9 / 13.7
5-15	45-55	5-15	5-15	2.5	1.55 / 2.81 / 2.33	87.4 / 94.6 / 96.6	0.1 / 0.3 / 1.4	10.7 / 2.9 / 6.6
5-15	65-75	5-15	5-15	2.5	1.69 / 2.91 / 2.44	88.3 / 93.0 / 95.1	0.0 / 0.8 / 1.6	11.2 / 1.7 / 4.8
5-15	85-95	5-15	5-15	2.5	1.66 / 2.87 / 2.42	86.8 / 94.0 / 96.7	0.0 / 1.5 / 2.2	5.8 / 0.5 / 2.1
5-15	25-35	5-15	25-35	2.5	1.54 / 3.04 / 2.52	78.9 / 93.5 / 96.1	0.5 / 1.2 / 1.3	1.2 / 0.7 / 1.6
5-15	45-55	5-15	25-35	2.5	1.59 / 2.87 / 2.50	70.7 / 93.0 / 95.8	1.5 / 5.1 / 3.6	0.3 / 0.1 / 0.4
5-15	65-75	5-15	25-35	2.5	1.68 / 2.73 / 2.45	70.9 / 93.8 / 96.4	2.8 / 7.9 / 6.0	0.0 / 0.0 / 0.0
5-15	85-95	5-15	25-35	2.5	1.82 / 2.74 / 2.50	75.9 / 93.2 / 94.9	6.2 / 16.4 / 13.8	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	2.5	1.70 / 2.71 / 2.48	58.7 / 94.1 / 96.0	6.5 / 15.1 / 11.4	0.0 / 0.0 / 0.0
5-15	65-75	5-15	45-55	2.5	1.83 / 2.67 / 2.49	65.3 / 93.7 / 95.7	13.2 / 25.3 / 21.9	0.0 / 0.0 / 0.1
5-15	85-95	5-15	45-55	2.5	1.98 / 2.66 / 2.54	74.1 / 94.9 / 95.7	28.9 / 44.7 / 42.0	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	2.5	1.99 / 2.63 / 2.51	73.2 / 93.8 / 95.3	30.6 / 39.3 / 38.0	0.0 / 0.0 / 0.1
5-15	85-95	5-15	65-75	2.5	2.13 / 2.56 / 2.49	85.0 / 94.6 / 94.8	51.3 / 57.6 / 57.6	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	2.5	2.35 / 2.55 / 2.53	92.8 / 94.5 / 95.0	78.5 / 79.8 / 80.4	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	2.5	1.24 / 2.80 / 2.41	64.1 / 97.1 / 97.0	0.0 / 0.0 / 1.2	3.7 / 6.9 / 15.7
25-35	25-35	5-15	5-15	2.5	1.36 / 2.97 / 2.43	61.9 / 95.0 / 95.8	0.0 / 0.3 / 1.2	2.2 / 0.3 / 3.9
25-35	45-55	5-15	5-15	2.5	1.47 / 3.07 / 2.51	62.7 / 94.1 / 96.9	0.5 / 2.2 / 2.6	1.4 / 0.2 / 0.6
25-35	65-75	5-15	5-15	2.5	1.56 / 3.05 / 2.49	65.4 / 92.6 / 94.4	0.5 / 5.9 / 8.0	1.2 / 0.0 / 0.5
25-35	85-95	5-15	5-15	2.5	1.67 / 2.97 / 2.49	70.4 / 92.5 / 94.5	0.8 / 9.7 / 13.3	1.1 / 0.1 / 0.2
25-35	25-35	5-15	25-35	2.5	1.43 / 2.75 / 2.49	24.4 / 94.7 / 96.2	0.8 / 4.7 / 6.1	0.0 / 0.0 / 0.1
25-35	45-55	5-15	25-35	2.5	1.55 / 2.72 / 2.54	26.3 / 94.8 / 95.7	3.5 / 16.2 / 16.2	0.0 / 0.0 / 0.0
25-35	65-75	5-15	25-35	2.5	1.66 / 2.71 / 2.49	36.0 / 92.7 / 95.2	11.0 / 34.4 / 30.3	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	2.5	1.81 / 2.73 / 2.52	47.6 / 92.7 / 95.6	25.7 / 57.9 / 55.8	0.0 / 0.0 / 0.0
25-35	45-55	5-15	45-55	2.5	1.63 / 2.55 / 2.50	14.8 / 94.9 / 95.7	24.4 / 50.1 / 47.7	0.0 / 0.0 / 0.0
25-35	65-75	5-15	45-55	2.5	1.78 / 2.58 / 2.53	25.8 / 94.1 / 93.8	54.1 / 78.1 / 75.4	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	2.5	1.93 / 2.56 / 2.50	43.4 / 93.6 / 95.1	74.1 / 88.6 / 87.5	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	2.5	1.91 / 2.52 / 2.50	37.4 / 95.3 / 96.0	82.5 / 90.7 / 89.6	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	2.5	2.09 / 2.51 / 2.49	63.1 / 95.5 / 95.9	95.0 / 97.7 / 97.5	0.0 / 0.0 / 0.0
25-35	85-95	5-15	85-95	2.5	2.30 / 2.50 / 2.52	85.6 / 95.0 / 95.6	99.6 / 99.6 / 99.6	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	2.5	1.22 / 2.43 / 2.18	73.0 / 97.5 / 96.1	0.0 / 0.0 / 0.1	6.6 / 4.8 / 15.4
45-55	25-35	5-15	5-15	2.5	1.37 / 2.89 / 2.52	53.9 / 93.0 / 96.2	0.1 / 0.5 / 0.5	0.6 / 0.4 / 3.8
45-55	45-55	5-15	5-15	2.5	1.44 / 2.91 / 2.51	54.4 / 92.8 / 95.0	0.0 / 2.1 / 3.6	0.6 / 0.1 / 0.8

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	65-75	5-15	5-15	2.5	1.56 / 2.96 / 2.51	59.8 / 93.0 / 94.7	0.3 / 5.3 / 10.0	0.4 / 0.0 / 0.1
45-55	85-95	5-15	5-15	2.5	1.66 / 2.98 / 2.49	68.0 / 92.1 / 95.3	0.1 / 10.8 / 20.9	0.5 / 0.0 / 0.1
45-55	25-35	5-15	25-35	2.5	1.40 / 2.55 / 2.50	15.4 / 95.6 / 96.6	0.6 / 4.1 / 6.7	0.0 / 0.0 / 0.1
45-55	45-55	5-15	25-35	2.5	1.51 / 2.58 / 2.50	17.6 / 94.0 / 94.4	3.5 / 18.2 / 18.9	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	2.5	1.63 / 2.61 / 2.50	25.4 / 92.6 / 93.2	12.0 / 40.1 / 41.7	0.0 / 0.0 / 0.0
45-55	85-95	5-15	25-35	2.5	1.78 / 2.68 / 2.52	41.9 / 93.2 / 94.9	24.1 / 62.3 / 65.8	0.0 / 0.0 / 0.0
45-55	45-55	5-15	45-55	2.5	1.58 / 2.44 / 2.53	8.2 / 93.9 / 94.4	25.6 / 54.4 / 55.1	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	2.5	1.75 / 2.48 / 2.51	20.1 / 93.7 / 94.4	59.5 / 81.8 / 83.4	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	2.5	1.90 / 2.51 / 2.50	36.8 / 95.1 / 95.6	81.6 / 93.8 / 94.8	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	2.5	1.88 / 2.45 / 2.52	27.9 / 93.9 / 95.6	88.4 / 94.3 / 94.2	0.0 / 0.0 / 0.0
45-55	85-95	5-15	65-75	2.5	2.07 / 2.48 / 2.49	55.3 / 94.7 / 95.4	97.7 / 98.9 / 98.9	0.0 / 0.0 / 0.1
45-55	85-95	5-15	85-95	2.5	2.27 / 2.47 / 2.51	79.6 / 94.3 / 94.8	99.9 / 99.9 / 99.9	0.0 / 0.0 / 0.1
25-35	5-15	25-35	5-15	2.5	1.12 / 2.26 / 2.36	0.4 / 95.7 / 93.2	0.1 / 0.0 / 7.4	0.0 / 0.2 / 10.2
25-35	25-35	25-35	5-15	2.5	1.20 / 2.28 / 2.47	0.0 / 94.6 / 93.6	0.4 / 0.6 / 14.1	0.0 / 0.0 / 1.8
25-35	45-55	25-35	5-15	2.5	1.24 / 2.30 / 2.52	0.0 / 93.0 / 94.0	0.9 / 3.9 / 23.8	0.0 / 0.0 / 0.6
25-35	65-75	25-35	5-15	2.5	1.30 / 2.33 / 2.51	0.0 / 94.0 / 94.2	2.2 / 10.6 / 35.2	0.0 / 0.0 / 0.1
25-35	85-95	25-35	5-15	2.5	1.36 / 2.35 / 2.51	0.0 / 92.6 / 94.7	6.7 / 22.5 / 56.3	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	2.5	1.33 / 2.26 / 2.53	0.1 / 92.6 / 94.2	4.0 / 13.7 / 34.2	0.0 / 0.0 / 0.1
25-35	45-55	25-35	25-35	2.5	1.40 / 2.25 / 2.54	0.0 / 90.6 / 94.2	16.1 / 45.9 / 61.9	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	2.5	1.50 / 2.29 / 2.51	0.0 / 90.4 / 95.4	43.6 / 73.3 / 83.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	2.5	1.59 / 2.32 / 2.51	0.6 / 90.2 / 94.3	69.5 / 90.6 / 95.4	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	2.5	1.51 / 2.25 / 2.51	0.0 / 87.9 / 93.7	56.0 / 84.3 / 87.8	0.0 / 0.0 / 0.0
25-35	65-75	25-35	45-55	2.5	1.64 / 2.30 / 2.52	0.1 / 89.4 / 94.9	91.8 / 97.7 / 98.6	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	2.5	1.78 / 2.36 / 2.53	1.7 / 91.9 / 94.6	99.1 / 99.8 / 100.0	0.0 / 0.0 / 0.1
25-35	65-75	25-35	65-75	2.5	1.80 / 2.36 / 2.53	2.6 / 89.8 / 95.1	99.5 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	65-75	2.5	1.99 / 2.41 / 2.53	19.3 / 92.5 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	2.5	2.24 / 2.45 / 2.51	69.9 / 93.5 / 94.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.1
45-55	5-15	25-35	5-15	2.5	1.12 / 1.99 / 2.34	11.7 / 94.8 / 90.9	0.0 / 0.1 / 5.6	0.0 / 0.2 / 14.5
45-55	25-35	25-35	5-15	2.5	1.17 / 2.09 / 2.51	0.0 / 91.8 / 92.2	0.1 / 0.6 / 9.3	0.0 / 0.0 / 1.5
45-55	45-55	25-35	5-15	2.5	1.22 / 2.15 / 2.53	0.0 / 92.4 / 92.8	0.7 / 2.6 / 21.9	0.0 / 0.0 / 0.2
45-55	65-75	25-35	5-15	2.5	1.27 / 2.18 / 2.52	0.0 / 90.9 / 94.5	1.0 / 6.5 / 37.7	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	2.5	1.34 / 2.26 / 2.49	0.1 / 92.7 / 93.5	4.0 / 13.7 / 55.4	0.0 / 0.0 / 0.0
45-55	25-35	25-35	25-35	2.5	1.27 / 2.03 / 2.53	0.0 / 86.0 / 92.7	1.1 / 6.6 / 26.2	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	2.5	1.35 / 2.10 / 2.50	0.0 / 84.3 / 93.6	9.9 / 29.3 / 52.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	2.5	1.45 / 2.18 / 2.51	0.0 / 86.1 / 93.8	30.1 / 59.2 / 80.7	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	2.5	1.57 / 2.25 / 2.51	0.9 / 86.9 / 94.3	56.8 / 83.3 / 94.8	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	2.5	1.45 / 2.14 / 2.52	0.0 / 79.8 / 94.5	45.7 / 72.4 / 84.8	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	2.5	1.59 / 2.19 / 2.50	0.1 / 79.4 / 94.3	81.6 / 94.7 / 97.9	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
45-55	85-95	25-35	45-55	2.5	1.74 / 2.27 / 2.50	1.3 / 86.1 / 95.7	97.3 / 99.7 / 99.8	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	2.5	1.75 / 2.27 / 2.51	1.2 / 83.8 / 94.6	98.7 / 99.7 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	2.5	1.97 / 2.36 / 2.52	17.4 / 90.7 / 95.8	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	2.5	2.21 / 2.42 / 2.51	66.2 / 94.4 / 95.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	2.5	1.11 / 1.58 / 2.09	42.0 / 91.9 / 95.6	0.0 / 0.0 / 1.6	0.0 / 0.3 / 19.0
45-55	25-35	45-55	5-15	2.5	1.15 / 1.76 / 2.29	6.5 / 87.1 / 95.5	0.0 / 0.2 / 4.3	0.0 / 0.0 / 8.2
45-55	45-55	45-55	5-15	2.5	1.20 / 1.84 / 2.38	4.8 / 85.2 / 92.3	0.0 / 0.3 / 12.2	0.0 / 0.0 / 4.3
45-55	65-75	45-55	5-15	2.5	1.27 / 1.92 / 2.38	7.0 / 84.6 / 93.4	0.4 / 1.8 / 23.7	0.0 / 0.0 / 0.8
45-55	85-95	45-55	5-15	2.5	1.33 / 1.99 / 2.43	9.4 / 87.3 / 94.6	0.4 / 3.9 / 35.4	0.0 / 0.0 / 0.0
45-55	25-35	45-55	25-35	2.5	1.22 / 1.83 / 2.49	0.0 / 75.4 / 94.2	0.4 / 2.4 / 18.9	0.0 / 0.0 / 0.2
45-55	45-55	45-55	25-35	2.5	1.28 / 1.88 / 2.48	0.0 / 70.0 / 95.2	2.4 / 9.8 / 35.8	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	2.5	1.37 / 1.97 / 2.49	0.0 / 71.9 / 95.4	12.5 / 32.2 / 61.4	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	2.5	1.48 / 2.07 / 2.49	0.1 / 78.7 / 95.9	28.3 / 55.4 / 81.1	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	2.5	1.38 / 1.97 / 2.50	0.0 / 66.3 / 95.0	21.4 / 42.7 / 68.9	0.0 / 0.0 / 0.0
45-55	65-75	45-55	45-55	2.5	1.51 / 2.07 / 2.50	0.1 / 72.1 / 94.8	56.7 / 78.8 / 91.1	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	2.5	1.65 / 2.16 / 2.51	0.5 / 79.1 / 94.3	85.2 / 93.6 / 97.2	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	2.5	1.68 / 2.16 / 2.51	0.9 / 76.1 / 95.5	91.5 / 97.6 / 99.2	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	2.5	1.89 / 2.26 / 2.51	13.1 / 83.9 / 95.7	99.4 / 99.9 / 99.9	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	2.5	2.17 / 2.38 / 2.51	62.0 / 91.7 / 93.1	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
5-15	5-15	5-15	5-15	3.0	2.38 / 2.60 / 2.71	95.9 / 99.3 / 93.1	0.0 / 0.0 / 2.3	58.3 / 23.7 / 27.8
5-15	25-35	5-15	5-15	3.0	1.80 / 3.42 / 2.85	88.7 / 97.3 / 95.1	0.4 / 0.1 / 2.8	21.0 / 11.0 / 13.4
5-15	45-55	5-15	5-15	3.0	1.69 / 3.65 / 2.86	79.5 / 94.3 / 94.4	0.1 / 1.1 / 3.4	11.8 / 3.1 / 5.1
5-15	65-75	5-15	5-15	3.0	1.75 / 3.61 / 2.88	79.1 / 93.4 / 95.3	0.1 / 2.9 / 2.8	8.6 / 1.3 / 3.1
5-15	85-95	5-15	5-15	3.0	1.84 / 3.58 / 2.98	79.0 / 93.2 / 95.4	1.0 / 4.8 / 5.3	5.1 / 0.3 / 1.5
5-15	25-35	5-15	25-35	3.0	1.64 / 3.87 / 3.07	69.6 / 92.5 / 95.2	0.6 / 3.1 / 4.2	1.4 / 1.2 / 2.5
5-15	45-55	5-15	25-35	3.0	1.78 / 3.57 / 3.05	56.3 / 92.3 / 94.2	3.0 / 12.5 / 9.7	0.1 / 0.2 / 0.3
5-15	65-75	5-15	25-35	3.0	1.87 / 3.38 / 3.00	56.2 / 93.3 / 94.9	6.8 / 21.0 / 18.8	0.0 / 0.0 / 0.2
5-15	85-95	5-15	25-35	3.0	2.01 / 3.29 / 2.98	59.1 / 91.7 / 94.1	15.0 / 34.4 / 31.2	0.0 / 0.0 / 0.0
5-15	45-55	5-15	45-55	3.0	1.92 / 3.33 / 3.00	43.7 / 92.6 / 95.4	13.7 / 32.3 / 31.6	0.0 / 0.0 / 0.1
5-15	65-75	5-15	45-55	3.0	2.06 / 3.26 / 3.03	52.4 / 93.4 / 95.4	31.6 / 48.1 / 45.4	0.0 / 0.0 / 0.0
5-15	85-95	5-15	45-55	3.0	2.23 / 3.14 / 3.02	58.0 / 94.0 / 95.9	52.7 / 71.1 / 71.2	0.0 / 0.0 / 0.0
5-15	65-75	5-15	65-75	3.0	2.26 / 3.15 / 3.01	62.5 / 93.5 / 95.2	54.8 / 65.7 / 65.6	0.0 / 0.0 / 0.3
5-15	85-95	5-15	65-75	3.0	2.47 / 3.08 / 3.01	74.9 / 94.1 / 95.2	78.5 / 84.5 / 84.5	0.0 / 0.0 / 0.0
5-15	85-95	5-15	85-95	3.0	2.76 / 3.05 / 3.04	88.7 / 94.8 / 95.7	94.0 / 94.5 / 94.7	0.0 / 0.0 / 0.0
25-35	5-15	5-15	5-15	3.0	1.24 / 2.94 / 2.59	45.5 / 97.7 / 94.5	0.0 / 0.0 / 3.0	2.3 / 8.6 / 13.5
25-35	25-35	5-15	5-15	3.0	1.41 / 3.56 / 2.91	43.1 / 94.5 / 93.9	0.0 / 1.1 / 3.8	1.8 / 1.4 / 3.2
25-35	45-55	5-15	5-15	3.0	1.55 / 3.56 / 2.97	47.3 / 93.3 / 95.1	0.0 / 4.4 / 8.0	0.5 / 0.4 / 0.4
25-35	65-75	5-15	5-15	3.0	1.66 / 3.57 / 2.96	50.3 / 92.6 / 94.9	1.2 / 10.9 / 17.2	0.5 / 0.4 / 0.1

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
25-35	85-95	5-15	5-15	3.0	1.81 / 3.57 / 2.95	56.0 / 91.5 / 95.9	1.0 / 20.1 / 31.9	0.8 / 0.0 / 0.0
25-35	25-35	5-15	25-35	3.0	1.49 / 3.10 / 2.90	10.1 / 94.3 / 95.7	1.1 / 11.0 / 12.6	0.0 / 0.0 / 0.4
25-35	45-55	5-15	25-35	3.0	1.66 / 3.15 / 3.01	13.0 / 93.0 / 95.9	10.2 / 37.7 / 42.4	0.0 / 0.0 / 0.1
25-35	65-75	5-15	25-35	3.0	1.79 / 3.17 / 3.00	19.1 / 92.3 / 96.3	22.5 / 63.1 / 64.7	0.0 / 0.0 / 0.0
25-35	85-95	5-15	25-35	3.0	1.97 / 3.17 / 3.00	29.4 / 92.2 / 95.0	42.9 / 83.8 / 85.2	0.0 / 0.0 / 0.1
25-35	45-55	5-15	45-55	3.0	1.73 / 2.86 / 2.94	3.9 / 92.3 / 96.2	46.0 / 74.1 / 75.1	0.0 / 0.0 / 0.1
25-35	65-75	5-15	45-55	3.0	1.94 / 2.94 / 2.98	9.8 / 93.4 / 96.4	78.9 / 94.5 / 94.0	0.0 / 0.0 / 0.0
25-35	85-95	5-15	45-55	3.0	2.12 / 2.96 / 3.00	22.2 / 93.1 / 95.8	93.2 / 99.3 / 98.9	0.0 / 0.0 / 0.0
25-35	65-75	5-15	65-75	3.0	2.12 / 2.92 / 3.01	20.1 / 90.6 / 93.6	95.9 / 98.5 / 98.4	0.0 / 0.0 / 0.0
25-35	85-95	5-15	65-75	3.0	2.37 / 2.95 / 3.00	41.0 / 92.6 / 96.3	99.6 / 99.7 / 99.6	0.0 / 0.0 / 0.1
25-35	85-95	5-15	85-95	3.0	2.66 / 2.95 / 3.02	73.6 / 93.6 / 95.5	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	5-15	5-15	3.0	1.29 / 2.77 / 2.52	63.3 / 97.6 / 95.1	0.0 / 0.0 / 0.3	5.9 / 4.9 / 10.0
45-55	25-35	5-15	5-15	3.0	1.40 / 3.25 / 2.88	38.0 / 93.3 / 95.8	0.0 / 0.6 / 2.6	0.3 / 1.1 / 1.1
45-55	45-55	5-15	5-15	3.0	1.50 / 3.40 / 2.96	40.3 / 92.7 / 95.4	0.1 / 5.1 / 12.5	0.4 / 0.3 / 0.5
45-55	65-75	5-15	5-15	3.0	1.59 / 3.29 / 2.88	41.5 / 94.3 / 94.3	0.2 / 8.5 / 24.0	0.3 / 0.0 / 0.3
45-55	85-95	5-15	5-15	3.0	1.77 / 3.51 / 2.98	52.7 / 92.5 / 94.5	0.9 / 20.6 / 46.0	0.0 / 0.1 / 0.0
45-55	25-35	5-15	25-35	3.0	1.44 / 2.81 / 2.93	4.6 / 92.6 / 95.6	0.9 / 9.0 / 19.1	0.0 / 0.0 / 0.1
45-55	45-55	5-15	25-35	3.0	1.56 / 2.86 / 2.95	4.1 / 91.1 / 96.1	6.4 / 33.7 / 48.1	0.0 / 0.0 / 0.0
45-55	65-75	5-15	25-35	3.0	1.73 / 2.94 / 2.97	10.2 / 92.3 / 94.8	20.2 / 64.5 / 75.9	0.0 / 0.0 / 0.1
45-55	85-95	5-15	25-35	3.0	1.89 / 3.02 / 2.99	19.6 / 92.1 / 96.8	38.8 / 86.3 / 92.5	0.0 / 0.0 / 0.1
45-55	45-55	5-15	45-55	3.0	1.66 / 2.69 / 2.98	1.6 / 88.4 / 95.2	43.9 / 76.0 / 85.4	0.0 / 0.0 / 0.0
45-55	65-75	5-15	45-55	3.0	1.86 / 2.78 / 3.01	5.1 / 88.5 / 95.4	80.7 / 95.9 / 97.8	0.0 / 0.0 / 0.0
45-55	85-95	5-15	45-55	3.0	2.07 / 2.82 / 2.98	16.4 / 90.9 / 94.3	95.5 / 99.1 / 99.6	0.0 / 0.0 / 0.0
45-55	65-75	5-15	65-75	3.0	2.03 / 2.76 / 2.99	9.9 / 87.6 / 93.9	97.3 / 99.3 / 99.7	0.0 / 0.0 / 0.2
45-55	85-95	5-15	65-75	3.0	2.32 / 2.88 / 3.00	33.5 / 92.6 / 94.2	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	5-15	85-95	3.0	2.59 / 2.89 / 3.00	62.7 / 92.1 / 95.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	5-15	25-35	5-15	3.0	1.13 / 2.41 / 2.70	0.1 / 95.2 / 97.0	0.0 / 0.1 / 13.5	0.0 / 0.1 / 3.3
25-35	25-35	25-35	5-15	3.0	1.20 / 2.34 / 2.89	0.0 / 88.7 / 95.8	0.2 / 1.0 / 30.6	0.0 / 0.0 / 0.3
25-35	45-55	25-35	5-15	3.0	1.25 / 2.34 / 2.96	0.0 / 85.4 / 95.5	0.8 / 6.0 / 49.8	0.0 / 0.0 / 0.3
25-35	65-75	25-35	5-15	3.0	1.31 / 2.41 / 2.95	0.0 / 85.8 / 95.4	3.7 / 15.5 / 66.3	0.0 / 0.0 / 0.0
25-35	85-95	25-35	5-15	3.0	1.37 / 2.44 / 2.97	0.0 / 82.9 / 95.5	8.1 / 26.3 / 82.1	0.0 / 0.0 / 0.0
25-35	25-35	25-35	25-35	3.0	1.34 / 2.31 / 2.97	0.0 / 78.6 / 95.5	4.8 / 18.7 / 59.3	0.0 / 0.0 / 0.0
25-35	45-55	25-35	25-35	3.0	1.44 / 2.34 / 3.00	0.0 / 73.7 / 95.9	23.6 / 52.7 / 84.7	0.0 / 0.0 / 0.0
25-35	65-75	25-35	25-35	3.0	1.54 / 2.42 / 3.02	0.0 / 71.3 / 96.1	55.3 / 84.4 / 95.7	0.0 / 0.0 / 0.0
25-35	85-95	25-35	25-35	3.0	1.65 / 2.45 / 3.00	0.0 / 69.1 / 94.9	81.2 / 95.9 / 99.1	0.0 / 0.0 / 0.0
25-35	45-55	25-35	45-55	3.0	1.55 / 2.36 / 2.99	0.0 / 61.2 / 96.0	69.0 / 91.8 / 98.0	0.0 / 0.0 / 0.0
25-35	65-75	25-35	45-55	3.0	1.70 / 2.46 / 3.02	0.0 / 61.7 / 93.3	95.6 / 99.4 / 99.7	0.0 / 0.0 / 0.0
25-35	85-95	25-35	45-55	3.0	1.87 / 2.54 / 3.02	0.1 / 66.5 / 93.6	99.8 / 100.0 / 100.0	0.0 / 0.0 / 0.0

continued on next page

Supplementary Table 2 – *continued from previous page*

SNP_1		SNP_2						
MAF(%)	R^2(%)	MAF(%)	R^2(%)	HR	Expected HR	Coverage (%)	Power(%)	Failure rate (%)
25-35	65-75	25-35	65-75	3.0	1.90 / 2.56 / 3.00	0.0 / 66.5 / 94.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.1
25-35	85-95	25-35	65-75	3.0	2.16 / 2.68 / 3.01	2.4 / 76.2 / 97.0	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
25-35	85-95	25-35	85-95	3.0	2.53 / 2.83 / 3.01	43.7 / 89.6 / 95.2	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	25-35	5-15	3.0	1.13 / 1.98 / 2.84	4.0 / 90.1 / 93.2	0.0 / 0.0 / 13.8	0.0 / 0.3 / 10.1
45-55	25-35	25-35	5-15	3.0	1.17 / 2.07 / 3.07	0.0 / 83.1 / 93.1	0.1 / 0.7 / 25.9	0.0 / 0.0 / 2.0
45-55	45-55	25-35	5-15	3.0	1.22 / 2.14 / 3.05	0.0 / 79.7 / 91.9	0.7 / 2.0 / 41.0	0.0 / 0.0 / 0.1
45-55	65-75	25-35	5-15	3.0	1.28 / 2.22 / 3.01	0.0 / 80.6 / 94.8	1.7 / 7.0 / 59.5	0.0 / 0.0 / 0.0
45-55	85-95	25-35	5-15	3.0	1.36 / 2.39 / 3.07	0.0 / 84.5 / 94.7	4.2 / 15.5 / 82.7	0.0 / 0.0 / 0.1
45-55	25-35	25-35	25-35	3.0	1.27 / 2.05 / 3.05	0.0 / 64.4 / 92.4	1.3 / 7.5 / 53.0	0.0 / 0.0 / 0.0
45-55	45-55	25-35	25-35	3.0	1.37 / 2.16 / 3.06	0.0 / 58.0 / 93.9	11.3 / 31.4 / 75.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	25-35	3.0	1.48 / 2.27 / 3.06	0.0 / 61.0 / 93.7	34.5 / 66.9 / 93.8	0.0 / 0.0 / 0.0
45-55	85-95	25-35	25-35	3.0	1.61 / 2.36 / 3.02	0.0 / 64.6 / 92.4	62.9 / 86.0 / 98.2	0.0 / 0.0 / 0.0
45-55	45-55	25-35	45-55	3.0	1.47 / 2.21 / 3.02	0.0 / 46.8 / 94.0	51.7 / 78.9 / 95.7	0.0 / 0.0 / 0.0
45-55	65-75	25-35	45-55	3.0	1.64 / 2.34 / 3.04	0.0 / 50.9 / 93.5	88.4 / 97.5 / 99.6	0.0 / 0.0 / 0.0
45-55	85-95	25-35	45-55	3.0	1.82 / 2.46 / 3.01	0.1 / 61.9 / 95.2	98.2 / 99.3 / 99.9	0.0 / 0.0 / 0.0
45-55	65-75	25-35	65-75	3.0	1.83 / 2.45 / 3.01	0.1 / 55.5 / 95.1	99.5 / 99.9 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	65-75	3.0	2.11 / 2.61 / 3.01	1.5 / 72.7 / 95.1	99.9 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	25-35	85-95	3.0	2.50 / 2.80 / 3.01	40.3 / 87.4 / 94.8	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	5-15	45-55	5-15	3.0	1.10 / 1.60 / 2.69	24.2 / 87.1 / 95.9	0.0 / 0.0 / 4.7	0.0 / 0.1 / 21.8
45-55	25-35	45-55	5-15	3.0	1.14 / 1.73 / 2.77	1.7 / 76.8 / 96.1	0.0 / 0.0 / 11.5	0.0 / 0.0 / 7.5
45-55	45-55	45-55	5-15	3.0	1.21 / 1.81 / 2.84	1.0 / 72.5 / 95.8	0.0 / 0.5 / 21.1	0.0 / 0.0 / 3.6
45-55	65-75	45-55	5-15	3.0	1.27 / 1.90 / 2.87	1.2 / 70.1 / 95.4	0.4 / 0.6 / 34.8	0.0 / 0.0 / 1.2
45-55	85-95	45-55	5-15	3.0	1.36 / 2.07 / 2.99	4.5 / 76.3 / 95.2	0.6 / 3.4 / 55.5	0.0 / 0.0 / 0.2
45-55	25-35	45-55	25-35	3.0	1.21 / 1.81 / 3.00	0.0 / 49.5 / 93.4	0.4 / 2.1 / 27.9	0.0 / 0.0 / 0.4
45-55	45-55	45-55	25-35	3.0	1.28 / 1.91 / 2.98	0.0 / 39.4 / 94.0	3.0 / 11.6 / 48.1	0.0 / 0.0 / 0.0
45-55	65-75	45-55	25-35	3.0	1.39 / 2.03 / 3.00	0.0 / 43.1 / 94.4	12.2 / 31.4 / 73.2	0.0 / 0.0 / 0.0
45-55	85-95	45-55	25-35	3.0	1.52 / 2.18 / 2.99	0.0 / 53.5 / 94.7	31.8 / 56.5 / 87.6	0.0 / 0.0 / 0.0
45-55	45-55	45-55	45-55	3.0	1.39 / 2.02 / 3.00	0.0 / 35.2 / 94.4	19.9 / 44.5 / 76.7	0.0 / 0.0 / 0.0
45-55	65-75	45-55	45-55	3.0	1.54 / 2.15 / 2.98	0.0 / 37.0 / 95.2	57.3 / 78.7 / 94.3	0.0 / 0.0 / 0.0
45-55	85-95	45-55	45-55	3.0	1.72 / 2.33 / 3.00	0.1 / 53.2 / 94.0	84.5 / 94.2 / 98.6	0.0 / 0.0 / 0.0
45-55	65-75	45-55	65-75	3.0	1.76 / 2.32 / 3.01	0.0 / 47.5 / 94.8	94.0 / 98.0 / 99.1	0.0 / 0.0 / 0.0
45-55	85-95	45-55	65-75	3.0	2.05 / 2.53 / 3.00	3.3 / 67.2 / 95.4	99.4 / 100.0 / 100.0	0.0 / 0.0 / 0.0
45-55	85-95	45-55	85-95	3.0	2.47 / 2.75 / 3.02	49.3 / 87.2 / 95.1	100.0 / 100.0 / 100.0	0.0 / 0.0 / 0.0

CNVassoc: Association analysis of CNV data

Isaac Subirana[1,2,3], Ramon Diaz-Uriarte[4], Gavin Lucas[2], Juan-Ramon Gonzalez[5,1]

March 29, 2011

[1]CIBER Epidemiology and Public Health (CIBERESP), Spain
[2]Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, IMIM, Hospital del Mar Research Institute, Spain
[3]Statistics Department, University of Barcelona, Spain
[4]Spanish National Cancer Centre (CNIO), Spain
[5]Center for Research in Environmental Epidemiology (CREAL), Spain

jrgonzalez@creal.cat http://www.creal.cat/jrgonzalez/software.htm

Contents

1 Introduction

CNVassoc allows users to perform association analysis between CNVs and disease incorporating uncertainty of CNV genotype. This document provides an overview on the usage of the CNVassoc package. For more detailed information on the model and assumption please refer to article [3] and its supplementary material. We illustrate how to analyze CNV data by using some real data sets. The first data set belongs to a case-control study where peak intensities from MLPA assays were obtained for two different genes. The second example corresponds to the Neve dataset [6] that is available at Bioconductor. The data consists of 50 CGH arrays of 1MB resolution for patients diagnosed with breast cancer. All datasets are available directly from the CNVassoc package. Finally, we show examples with Poisson and Weibull-distributed phenotypes

Start by loading the package CNVassoc:

```
> library(CNVassoc)
```

and some required libraries

```
> library(xtable)
```

2 CNV from a single probe

2.1 The data

In order to illustrate how to assess association between CNV and disease, we use a data set including 360 cases and 291 controls. Data is to be published soon as described in [3]. The data contains peaks intensities for two genes arising from an MLPA assay. Note that Illumina or Affymetrix data, where log2 ratios are available instead of peak intensities, can be analyzed in the same way as we are illustrating.

The MLPA data set contains case control status as well as two simulated covariates (quanti and cov) that have been generated for illustrative purposes (e.g., association between a quantitative trait and CNV or how to adjust for covariates). To load the MLPA data just type

```
> data(dataMLPA)
> head(dataMLPA)
```

```
    id casco Gene1     Gene2 PCR.Gene1 PCR.Gene2 quanti   cov
1 H238     1  0.51 0.5385080        wt        wt  -0.61 10.83
2 H238     1  0.45 0.6392029        wt        wt  -0.13 10.69
3 H239     1  0.00 0.4831572       del        wt  -0.57  9.63
4 H239     1  0.00 0.4640072       del        wt  -1.40  9.87
5 H276     1  0.00 0.0000000       del       del   0.83 10.25
6 H276     1  0.00 0.0000000       del       del  -2.07 10.40
```

First, we look at the distribution of peak intensities for each of the two genes analyzed: see Figure 1.

Figure 1 shows the signals for Gene 1 and Gene 2. For both genes it is clear that there are 3 clusters corresponding to 0, 1 and 2 copies. However, the three peaks for Gene 2 are not so well separated as those of Gene 1 (the underlying distributions overlap much more). This fact leads to more uncertainty when inferring the copy number status for each individual. This will be illustrated in the next section.

In the CNVassoc package, a function called plotSignal has been implemented to plot the peak intensities for a gene. To illustrate this, a plot of the intensities of Gene 2 for each individual, distinguishing between cases and controls, can be performed by typing (see figure 2)

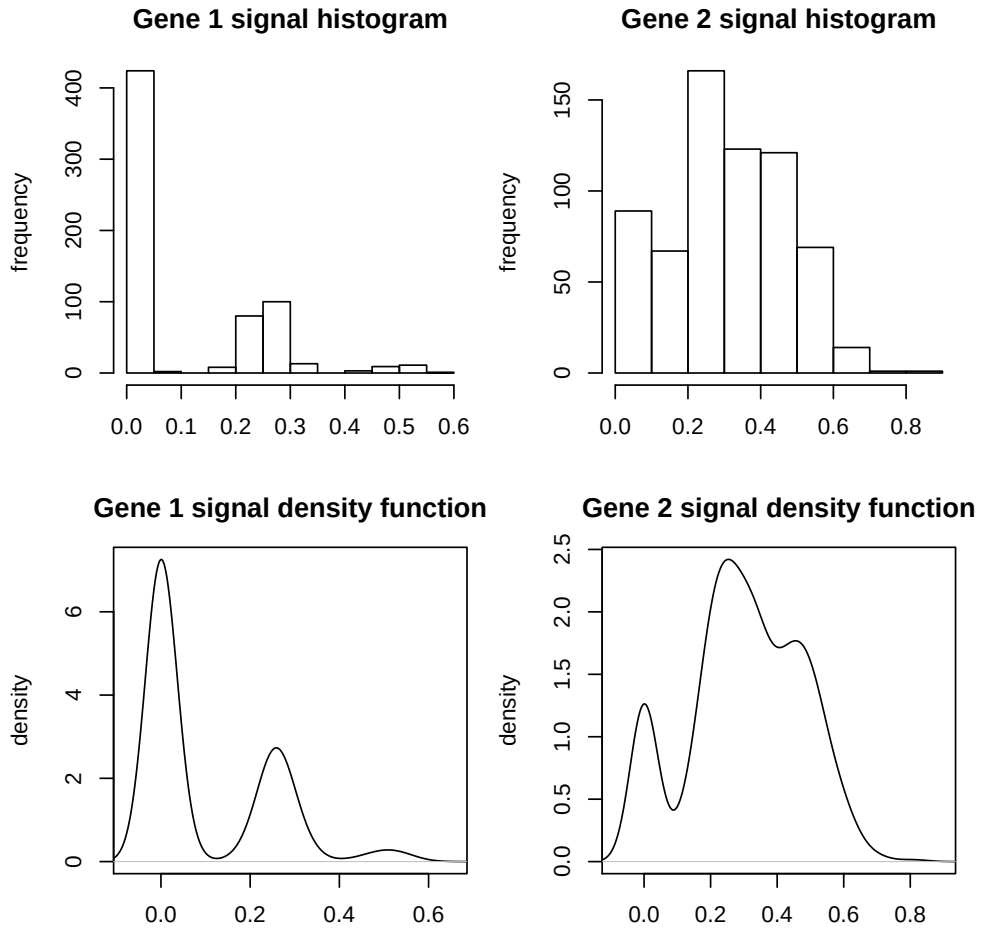


Figure 1: Signal distributions for Gene 1 and Gene 2

```
> plotSignal(dataMLPA$Gene2, case.control = dataMLPA$casco)
```

or, similarly but correlating the peak intensities with a quantitative phenotype (see figure 3) type

```
> plotSignal(dataMLPA$Gene2, case.control = dataMLPA$quanti)
```

In figure 3, the quantitative phenotype is plotted on the x-axis, instead of distinguishing points by shape, as in figure 2.

Also, it is possible to specify the number of cutoff points and place them interactively via `locator` on the previous plot, in order to infer the copy number status in a naive way. (More sophisticated ways of inferring copy number status will be dealt with in subsequent sections). To place 2 cutoff points, thereby defining 3 copy number status values or clusters, (note use of argument `n=2`) and store them as `cutpoints`:

```
\dontrun{
cutpoints<-plotSignal(dataMLPA$Gene2,case.control=dataMLPA$casco,n=2)
}
```

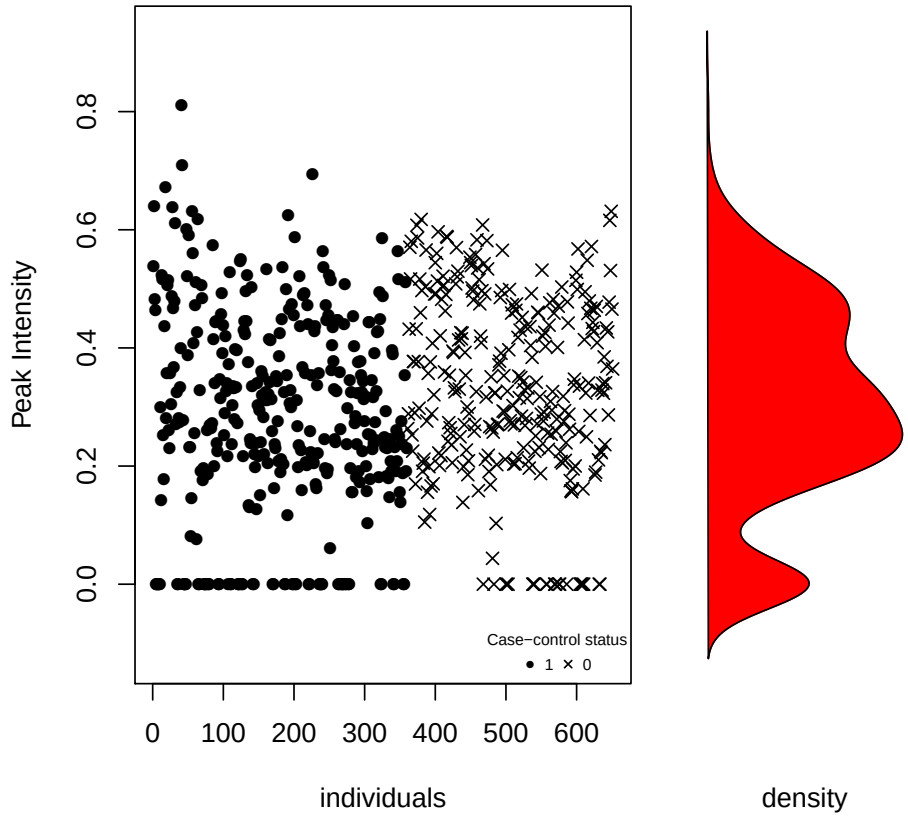


Figure 2: Signal distribution for Gene 2 using `plotSignal`

The plot generated in figure 4, is similar to that of 2, but using colours to distinguish copy number status values inferred from the cutoff points.

In this example, the cutoff points have been placed at:

```
> cutpoints

[1] 0.08470221 0.40485249
```

These stored cutoff points will be used in the following sections.

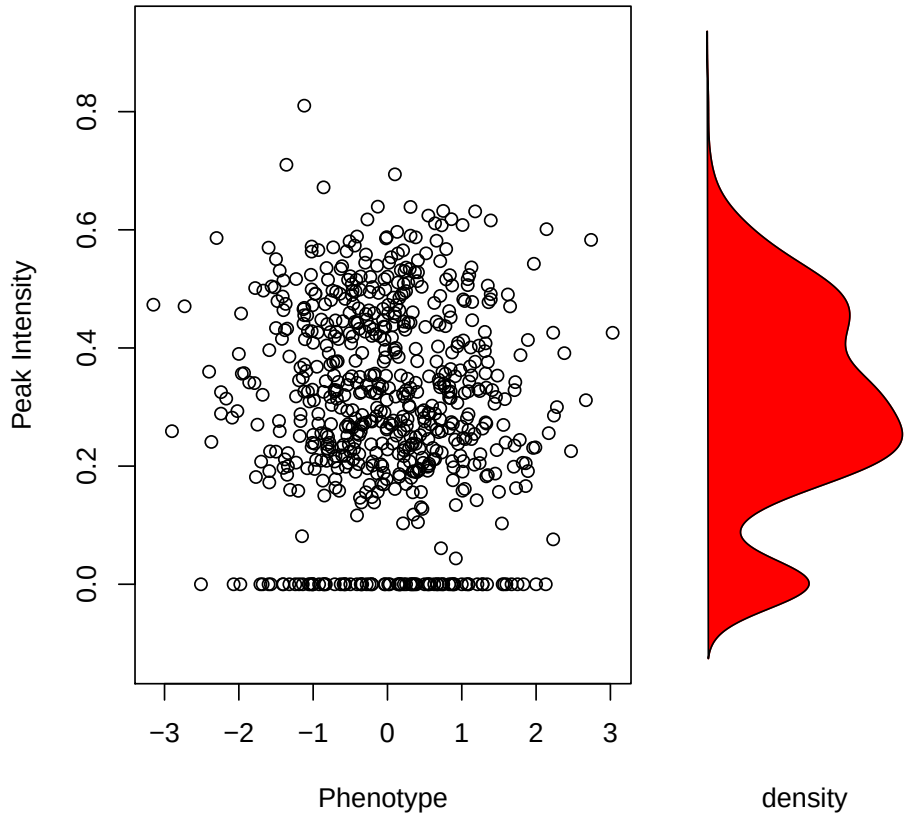


Figure 3: Signal distribution for Gene 2 using `plotSignal`

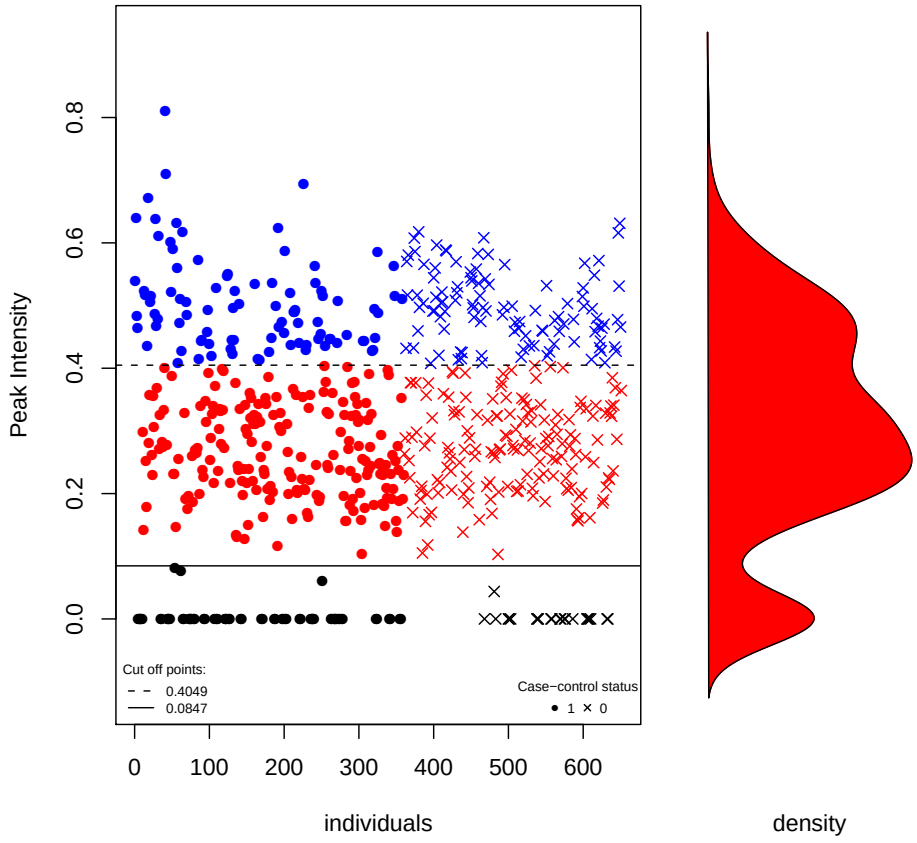


Figure 4: Signal distribution for Gene 2 using `plotSignal` once cutoff points have been set with `locator`

2.2 Inferring copy number status from signal data

2.2.1 From univariate signal intensity

The `cnv` function is used to infer the copy number status for each subject using the quantitative signal for an individual probe. This signal can be obtained from any platform (MLPA, Illumina, ...).

This function assumes a normal mixture model as other authors have proposed in the context of aCGH [7, 9]. It should be pointed out that in some instances, the intensity distributions (see Gene 1 in Figure 1) for a null allele are expected to be equal to 0. Due to experimental noise these intensities can deviate slightly from this theoretical value. For these cases, the normal mixture model fails because the underlying distribution of individuals with 0 copies is not normal. In these situations we fit a modified mixture model (see [3] for further details).

Figure 1 presents two distinctly different scenarios. For Gene 1 there are clearly three different status values, but for Gene 2 the situation is not so clear.

Function `cnv` provides various arguments to cope with all these issues. The calling for Gene 1 can be done by executing

```
> CNV.1 <- cnv(x = dataMLPA$Gene1, threshold.0 = 0.06, num.class = 3,
+     mix.method = "mixdist")
```

The argument `threshold.0=0.06` indicates that individuals with peak intensities lower than 0.06 will have 0 copies. Since there are three underlying copy number status values, we set argument `num.class` to 3. Argument `mix.method` indicates what algorithm to use in estimating the normal mixture model. `"mixdist"` uses a combination of a Newton-type method and the EM algorithm implemented in the `mixdist` library, while `"mclust"` uses the EM algotithm implemented in the `Mclust` library.

When the exact number of components for the mixture model is unknown (which may be the case for Gene 2), the function uses the Bayesian Information Criteria (BIC) to select the number of components. This is performed when the argument `num.class` is missing. In this case the function estimates the mixture model admitting from 2 up to 6 copy number status values.

```
> CNV.2 <- cnv(x = dataMLPA$Gene2, threshold.0 = 0.01, mix.method = "mixdist")
```

As we can see, the best model has a copy number status of 3. This result, obtained by using BIC, is as expected because we already know that this gene has 0, 1 and 2 copies (see [3]).

2.2.2 From other algorithms

The result of applying function `cnv` is an object of class `cnv` that, among other things, contains the posterior probabilities matrix for each individual. This information is then used in the association analysis where the uncertainty is taken into account. Posterior probabilities from any other calling algorithms can also be encapsulated in a `cnv` object to be further used in the analysis.

To illustrate this, we will use the posterior probability matrix that has been computed when inferring copy number for Gene 2 by using the normal mixture model. This information is saved as an attribute for an object of class `cnv`. A function called `getProbs` has been implemented to simplify accessing this attribute. Thus the probability matrix can be saved in an object `probs.2` like this:

```
> probs.2 <- getProbs(CNV.2)
```

Imagine that `probs.2` contains posterior probabilities obtained from some calling algorithm such as CANARY (from PLINK) or `GCHcall` (this will be further illustrated in Section 3). In this case, we create the object of class `cnv` that will be used in the association step by typing

```
> CNV.2probs <- cnv(probs.2)
```

2.2.3 From predetermined thresholds

Inferring copy number status for Gene 2 from previously specified threshold points (stored in vector `cutpoints`) can be done using the same `cnv` function but setting the argument `cutoffs` to `cutpoints`.

```
> CNV.2th <- cnv(x = dataMLPA$Gene2, cutoffs = cutpoints)
```

Now, the inferred copy number object `CNV.2th` contains the same information as it would if it had been created directly from probabilities.

2.3 Summarizing information

We have implemented two generic functions for an object of class `cnv`. The generic `print` function gives the results on inferred copy number status. It includes the means, variances and proportions of copy number clusters as well as the p value corresponding to the goodness-of-fit test for the selected number of classes.

```
> CNV.1

Inferred copy number variant by a quantitative signal
   Method: function mix {package: mixdist}

-. Number of individuals: 651
-. Copies 0, 1, 2
-. Estimated means: 0, 0.2543, 0.4958
-. Estimated variances: 0, 9e-04, 0.0012
-. Estimated proportions: 0.6544, 0.3088, 0.0369
-. Goodness-of-fit test: p-value= 0.6615318
```

and for Gene 2

```
> CNV.2

Inferred copy number variant by a quantitative signal
   Method: function mix {package: mixdist}

-. Number of individuals: 651
-. Copies 0, 1, 2
-. Estimated means: 0, 0.2435, 0.4469
-. Estimated variances: 0, 0.0041, 0.0095
-. Estimated proportions: 0.1306, 0.4187, 0.4507
-. Goodness-of-fit test: p-value= 0.4887659


-. Note: number of classes has been selected using the best BIC
```

This report differs slightly when the object was created from only posterior probabilities:

```
> CNV.2probs
```

```
-. Copy number variant
   Input data: called probabilities
-. Number of individuals: 651
-. Copies 0, 1, 2
-. Estimated proportions: 0.1306, 0.4187, 0.4507
```

Figure 5 shows the result of invoking the generic `plot` function on these objects.

```
> pdf("./figures/fig2a.pdf")
> plot(CNV.1, case.control = dataMLPA$casco, main = "Gene 1")
> dev.off()

windows
      2

> pdf("./figures/fig2b.pdf")
> plot(CNV.2, case.control = dataMLPA$casco, main = "Gene 2")
> dev.off()

windows
      2
```

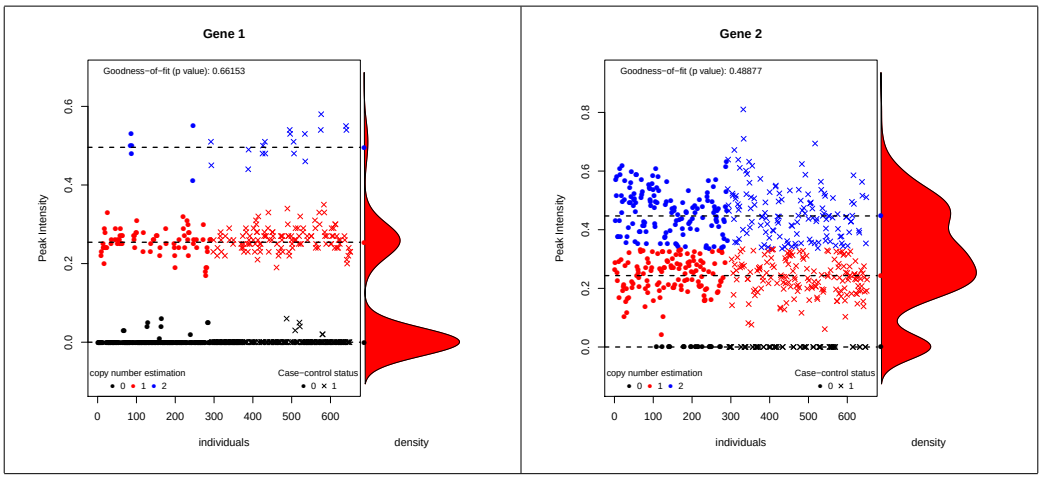


Figure 5: Signal distribution by case control, and inferred number of copies

In figure 5 the signal is coloured by the inferred (most probable) copy number, while cases and controls are distinguished by shape. This last option is specified by the argument `case.control`. On the right side of the plot, a density function of signal distribution is drawn. The p-value of goodness-of-fit test is the same as this described in the beginning of this section. It indicates whether the assumed normal mixture model (with a given number of components) is correct or not. Notice that for both genes the intensity data fits our the model well (goodness-of-fit p-values > 0.1).

The action of `plot` when only posterior probabilities are available gives a different result (Figure 6). Two barplots are created for cases and controls (when argument `case.control` is used). Both are split by the copy number frequency.

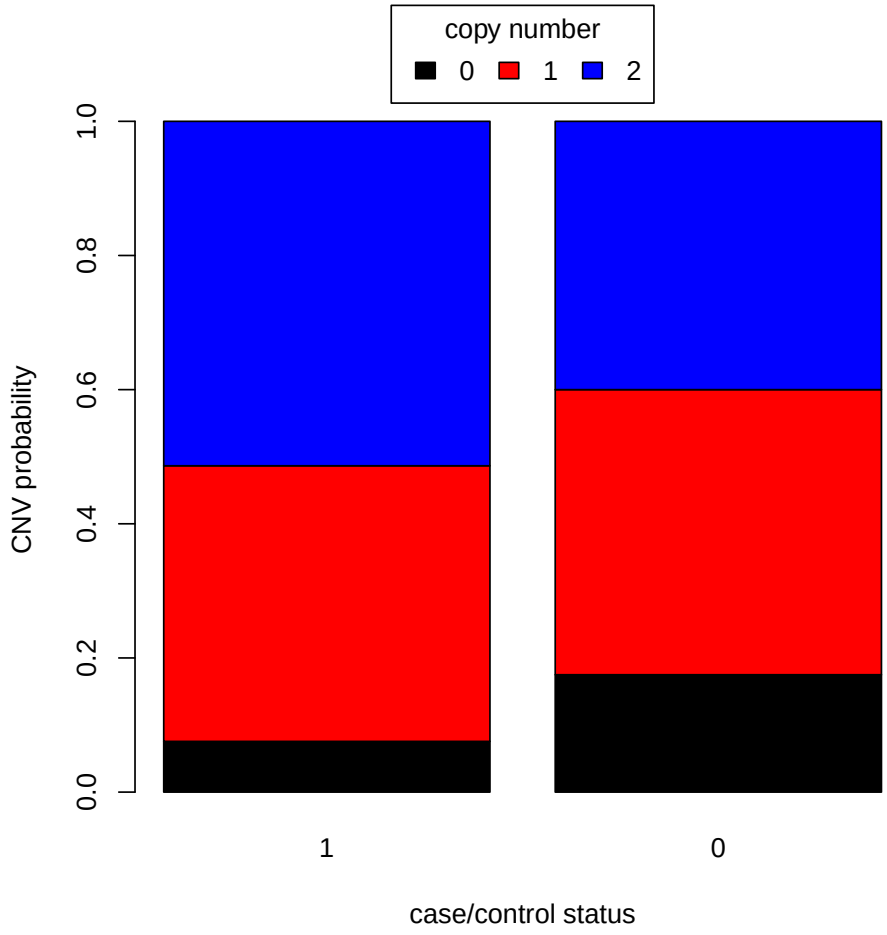


Figure 6: Estimated copy number frequencies for Gene 1 and Gene 2

2.4 Measuring uncertainty in inferring copy number status

The function `getQualityScore` uses information from an object of class `cnv` to compute a value that indicates how much the underlying copy number distribution (peak intensities) are mixed or overlapped. The more separated these peaks are (less uncertainty), the larger the quality score is.

Three measures of uncertainty are currently implemented. The first one is the same as that defined in the `CNVtools` package, the second is the estimated probabilty of good classification (PGC), and the third is defined as the the proportion of individuals with a confidence score (described in [4]) bigger than 0.1.
To choose PGC method type

```
> CNVassoc::getQualityScore(CNV.1, type = "class")

--Probability of good classification: 0.9999963

> CNVassoc::getQualityScore(CNV.2, type = "class")

--Probability of good classification: 0.9096771
```

To choose the measure defined in the `CNVtools` package:

```
> CNVassoc::getQualityScore(CNV.1, type = "CNVtools")

--CNVtools Quality Score: 25.16849

> CNVassoc::getQualityScore(CNV.2, type = "CNVtools")

--CNVtools Quality Score: 3.057171
```

And to choose the third measure:

```
> CNVassoc::getQualityScore(CNV.1, type = "CANARY")

--Probability to have a 'CANARY confidence index' > 0.1 : 0

> CNVassoc::getQualityScore(CNV.2, type = "CANARY")

--Probability to have a 'CANARY confidence index' > 0.1 : 0.3024652
```

It is clear that in Gene 1 there is much less uncertainty, because the PGC is greater than 99%, the measure of `CNVtools` package is higher than 25 (CNVtools recommends a quality score of 4 or larger), or the "CANARY" measure is almost 0. This fact can also be seen in Figure 5 where the underlying distributions of signal intensity are very well separated. On the other hand, the PGC for Gene 2 is 91.3%, and the `CNVtools` package value is about 3 indicating that more uncertainty is present, and the "CANARY" type measure for Gene 2 tells that up to 30% of individuals have a poor confidence score. When cnv object has been created directly from probabilities (obtained from any other calling algorithm), only `type="CANARY"` method can be computed. In [5], it is suggested that, when proportion of individuals with confidence score > 0.1 is greater than 10%, this particular CNV should be removed from the analysis under a best-guess strategy in performing the association test.

2.5 Assessing associations between CNV and disease

The function `CNVassoc` carries out association analysis between CNV and disease. This function incorporates calling uncertainty by using a latent class model as described in [3]. The function can analyze both binary and quantitative traits. In the first case, a linear regression is performed, and, in the second, a logistic regression. The regression model can be selected by using the argument `case.control`. Nonetheless, the program automatically detects whether or not a quantitative trait is being analyzed so it need not be specified.

The function also allows the user to fit a model with additive or multiplicative effects of CNV. This can be set through the argument `model`. Possible values are "add" for an additive effect or "mul" for a multiplicative effect.

The function `CNVassoc` returns an object of class `CNVassoc`. This class of object has some properties in common with objects of class `glm`, such as `coef` or `summary` among others.

2.5.1 Modelling association

The effect of a given CNV on case/control status (`casco` variable) can be fitted by typing

```
> model1mul <- CNVassoc(casco ~ CNV.1, data = dataMLPA, model = "mul")
> model2mul <- CNVassoc(casco ~ CNV.2, data = dataMLPA, model = "mul")
```

By default, a short summary is printed (similar to `glm` objects)

```
> model1mul

Call:  CNVassoc(formula = casco ~ CNV.1, data = dataMLPA, model = "mul")

Coefficients:
         CNV0       CNV1       CNV2
CNVmult  0.0281709  0.5187566  1.0989109

Number of individuals: 651
Number of estimated parameters: 3
Deviance: 883.03

> model2mul

Call:  CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "mul")

Coefficients:
         CNV0        CNV1        CNV2
CNVmult   1.0520923   0.3122567  -0.0970782

Number of individuals: 651
Number of estimated parameters: 3
Deviance: 876.396
```

Note that the coefficients are a matrix with one row per variable and a column for each distinct copy number status. In this model, because there are no covariates and the CNV has a multiplicative effect, there is just one row (one intercept) and this is different among columns (copy number status).

By using the generic function `summary` we can obtain a more exhaustive output. In particular the odds ratio and its confidence intervals are printed as well as its p-value.

```
> summary(model1mul)

Call:
CNVassoc(formula = casco ~ CNV.1, data = dataMLPA, model = "mul")

Deviance: 883.0297
Number of parameters: 3
Number of individuals: 651

Coefficients:
         OR lower.lim upper.lim     SE   stat pvalue
CNV0 1.0000
CNV1 1.6333    1.1588    2.3020 0.1751 2.8017  0.005
CNV2 2.9175    1.1359    7.4937 0.4813 2.2247  0.026

(Dispersion parameter for  binomial  family taken to be  1 )


Covariance between coefficients:
     CNV0   CNV1   CNV2
CNV0 0.0094 0.0000 0.0000
CNV1        0.0213 0.0000
CNV2               0.2223

> summary(model2mul)

Call:
CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "mul")

Deviance: 876.396
Number of parameters: 3
Number of individuals: 651

Coefficients:
          OR lower.lim upper.lim      SE    stat   pvalue
CNV0  1.0000
CNV1  0.4772    0.2742    0.8304  0.2827 -2.6172    0.009
CNV2  0.3169    0.1834    0.5477  0.2791 -4.1169 3.84e-05

(Dispersion parameter for  binomial  family taken to be  1 )


Covariance between coefficients:
     CNV0    CNV1    CNV2
CNV0  0.0613  0.0000  0.0000
CNV1          0.0186 -0.0032
CNV2                  0.0166
```

By default, `CNVassoc` treats the response variable as a binary phenotype coded as 0/1. Since `CNVassoc` can handle other distributions such as Poisson or Weibull, the `family` argument must be

specified when the response is not distributed as a bernoulli. For instance, to deal with a normally distributed response variable, specify `family="gaussian"`

The following example presents the case of analyzing a quantitative normally distributed trait and adjusting the association by other covariates:

```
> mod <- CNVassoc(quanti ~ CNV.2 + cov, family = "gaussian", data = dataMLPA,
+     model = "add", emsteps = 10)
> mod

Call:  CNVassoc(formula = quanti ~ CNV.2 + cov, data = dataMLPA, model = "add",      family =

Coefficients:
           CNV0        CNV1        CNV2
intercept -0.1403761 -0.1403761 -0.1403761
CNVadd    -0.0792367 -0.0792367 -0.0792367
cov        0.0241877  0.0241877  0.0241877

Number of individuals: 651
Number of estimated parameters: 4
Deviance: 1824.57
```

Notice that in this case, we use new argument called `emsteps`. This is necessary for computational reasons. Initially performing some preliminary steps using the EM algorithm makes it easier to maximize the likelihood function using the Newton-Raphson procedure. In general, it is enough to perform a few iterations (no more than 10). As usual, the model is then summarized by typing

```
> summary(mod)

Call:
CNVassoc(formula = quanti ~ CNV.2 + cov, data = dataMLPA, model = "add",
    family = "gaussian", emsteps = 10)

Deviance: 1824.573
Number of parameters: 4
Number of individuals: 651

Coefficients:
               beta lower.lim upper.lim       SE     stat pvalue
(Intercept) -0.14038  -0.90687   0.62612  0.39108 -0.35895  0.720
trend       -0.07924  -0.19714   0.03866  0.06015 -1.31722  0.188
cov          0.02419  -0.05068   0.09906  0.03820  0.63321  0.527

(Dispersion parameter estimation for  gaussian family is 0.9650261 )


Covariance between coefficients:
          intercept CNVadd  cov
intercept  0.1529   -0.0041 -0.0146
CNVadd              0.0036 -0.0001
cov                        0.0015
```

Remember that for quantitative traits we obtain mean differences instead of odds ratios.

2.5.2 Testing associations

In the previous analysis we obtained p values corresponding to the comparison between every copy number status versus the reference (zero copies). Nonetheless, we are normally interested in testing the overall effect of CNV on disease. We have implemented the Wald test and the likelihood ratio test (LRT) to perform such omnibus testing. Both are available through the function `CNVtest` which requires an object of class `CNVassoc` as the input. To specify the type of test, set the argument `type` to `"Wald"` or `"LRT"`, respectively. For Gene 1,

```
> CNVtest(model1mul, type = "Wald")

----CNV Wald test----
Chi= 11.55332 (df= 2 ) , pvalue= 0.003099052

> CNVtest(model1mul, type = "LRT")

----CNV Likelihood Ratio Test----
Chi= 12.12081 (df= 2 ) , pvalue= 0.002333458
```

and for Gene 2,

```
> CNVtest(model2mul, type = "Wald")

----CNV Wald test----
Chi= 17.32966 (df= 2 ) , pvalue= 0.0001725492

> CNVtest(model2mul, type = "LRT")

----CNV Likelihood Ratio Test----
Chi= 18.75453 (df= 2 ) , pvalue= 8.462633e-05
```

Other generic functions like `logLik`, `coef`, `summary` or `update` can be applied to an object of class `CNVassoc` to get more information.

For a multiplicative CNV effect model and for a binary traits, it is possible to change the reference category of copy number status. This can be done by using the argument `ref` when executing the `summary` function. For example, if we want to one copy as the reference category just type:

```
> coef(summary(model1mul, ref = 2))

            OR lower.lim upper.lim        SE      stat      pvalue
CNV1 1.0000000        NA        NA        NA        NA          NA
CNV0 0.6122677 0.4344016 0.8629612 0.1751053 -2.801661 0.005084028
CNV2 1.7863140 0.6790498 4.6990928 0.4934862  1.175624 0.239745087
```

The same kind of results can be obtained if we assume an additive effect of CNV on the trait. In this case we need to set the `model` argument to `"add"`

```
> model2add <- CNVassoc(casco ~ CNV.2, data = dataMLPA, model = "add")
> model2add
```

```
Call:  CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "add")

Coefficients:
           CNV0       CNV1       CNV2
intercept   0.932028   0.932028   0.932028
CNVadd     -0.537731  -0.537731  -0.537731

Number of individuals: 651
Number of estimated parameters: 2
Deviance: 877.061
```

Notice that under an additive CNV effect the structure of coefficients are different from the multiplicative CNV effect. Now there are two rows, one for intercept and the other one for the slope (change of risk in increasing by one copy). These two values remain constant for every column (copy number status).

```
> summary(model2add)

Call:
CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "add")

Deviance: 877.0606
Number of parameters: 2
Number of individuals: 651

Coefficients:
           OR lower.lim upper.lim      SE    stat   pvalue
trend  0.5841    0.4530    0.7530  0.1296 -4.1477 3.36e-05

(Dispersion parameter for  binomial  family taken to be  1 )


Covariance between coefficients:
          intercept CNVadd
intercept  0.0374   -0.0228
CNVadd              0.0168
```

Finally, one might be interested in testing the additive effect. To do this, one can compare both additive and multiplicative models. It is straightforward to see that the additive model is a particular case of the multiplicative one, and therefore the first is nested in the second one.

To compare two nested models we use the generic function `anova` (NOTE: it is only implemented for comparing two models, both fitted with the `CNVassoc` function).

```
> anova(model2mul, model2add)

--- Likelihood ratio test comparing 2 CNVassoc models:


Model 1 call:  CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "mul")

Model 2 call:  CNVassoc(formula = casco ~ CNV.2, data = dataMLPA, model = "add")
```

```
Chi= 0.6645798 (df= 1 )  p-value= 0.4149477

  Note: the 2 models must be nested, and this function doesn't check this!
```

The likelihood ratio test is performed. In this case the p-value is not significant, indicating that an additive CNV effect can be assumed. In any case, one should consider the power of this test before making conclusions.

2.6 Analysing other genetic models

In assessing copy number variant effect on a disease `CNVassoc` package can deal with additive or multiplicative (see more details in [3]). The first one ('additive') suppose an equal increase in logit of risk (for case-control studies) or in mean (for a quantitative traits) for example, while 'multiplicative' makes no assumptions on CNV effect.

In the particular case that CNV has 3 categories ('copy lose', 'normal' or 'copy gain'), it may be useful to assume CNV effect not being additive or multiplicative, but dominant or recessive. That is, to compare the effect of 'copy lose' vs. the other two if a dominant effect is assumed, or to compare 'copy gain' vs. the other two if a recessive effect is assumed.

It is possible to assesse such 'dominant' or 'recessive' effect using `CNVassoc` package functions. To do so, few simple steps have to be done before performing the associations analysis. Here, we illustrate the required instructions to perform the association analysis assuming a recessive or a dominant effect, taking MLPA example data already present in the `CNVassoc` package.

a) **Package and data loading:** First, `CNVassoc` package and MLPA data are loaded

```
> library(CNVassoc)
> data(dataMLPA)
```

b) **Inferring copy number status:** Then, CNV from Gene2 signal intensities is inferred. And its copy number probabilities matrix is stored in 'probs' object.

```
> CNV <- cnv(x = dataMLPA$Gene2, threshold.0 = 0.01, mix.method = "mixdist")
> CNV

Inferred copy number variant by a quantitative signal
   Method: function mix {package: mixdist}

-. Number of individuals: 651
-. Copies 0, 1, 2
-. Estimated means: 0, 0.2435, 0.4469
-. Estimated variances: 0, 0.0041, 0.0095
-. Estimated proportions: 0.1306, 0.4187, 0.4507
-. Goodness-of-fit test: p-value= 0.4887659


-. Note: number of classes has been selected using the best BIC

> probs <- attr(CNV, "probabilities")
```

c) **Updating CNV to dominant or recessive:** Once CNV is inferred, some previous modifications to CNV object have to be done: For assessing the recessive effect, first and second copy number status has to be joined. To do so, first and second columns of 'probs' are added. Also, copy number status labels must be modified from 'CNVrec' and 'CNVdom' objects. In both cases, they can be set to 0,1:

```
> probsrec <- cbind(rowSums(probs[, 1:2]), probs[, 3])
> CNVrec <- cnv(probsrec, num.copies = c(0, 1))
> CNVrec

-. Copy number variant
   Input data: called probabilities
-. Number of individuals: 651
-. Copies 0, 1
-. Estimated proportions: 0.5493, 0.4507
```

And for assessing the dominant effect, we proceed the same way but adding the second and third columns:

```
> probsdom <- cbind(probs[, 1], rowSums(probs[, 2:3]))
> CNVdom <- cnv(probsdom, num.copies = c(0, 1))
> CNVdom

-. Copy number variant
   Input data: called probabilities
-. Number of individuals: 651
-. Copies 0, 1
-. Estimated proportions: 0.1306, 0.8694
```

d) **Performing association test:** Finally, association analysis is performed as usual, specifying a 'multiplicative' effect in the 'model' argument (note that an 'additive' effect could be set and the same results would be obtained). In this example, the Odds Ratio of category labelled as '1' vs '0' is displayed. When assessing the dominant model effect, '1' will contain 'copy-gain' and 'normal' categories, while '0' will contain 'copy-lose'. On the other hand, when assessing the recessive model effect, '1' will contain 'copy-gain', and '0' will contain 'normal' and 'copy-lose' categories.

The results of the association test are:

- for recessive

```
> summary(CNVassoc(casco ~ CNVrec, data = dataMLPA))

Call:
CNVassoc(formula = casco ~ CNVrec, data = dataMLPA)

Deviance: 883.644
Number of parameters: 2
Number of individuals: 651

Coefficients:
          OR lower.lim upper.lim      SE    stat pvalue
CNV0  1.0000
CNV1  0.5309    0.3672    0.7677  0.1882 -3.3650  0.001
```

```
(Dispersion parameter for  binomial  family taken to be  1 )


Covariance between coefficients:
     CNV0    CNV1
CNV0  0.0141 -0.0024
CNV1          0.0165
```

- and for dominant

```
> summary(CNVassoc(casco ~ CNVdom, data = dataMLPA))
Call:
CNVassoc(formula = casco ~ CNVdom, data = dataMLPA)

Deviance: 880.4668
Number of parameters: 2
Number of individuals: 651

Coefficients:
          OR lower.lim upper.lim      SE    stat   pvalue
CNV0  1.0000
CNV1  0.3856    0.2309    0.6438  0.2616 -3.6438 0.000269

(Dispersion parameter for  binomial  family taken to be  1 )


Covariance between coefficients:
     CNV0   CNV1
CNV0 0.0613 0.0000
CNV1        0.0071
```

3 CNV from aCGH

The analysis of aCGH data requires taking additional steps into account, due to the dependency across probes and the fact that CNVs are not measured with a unique probe. Table 1 shows four steps we recommend for the analysis of this kind of data. First, posterior probabilities should be obtained with an algorithm that considers probe correlation. We use, in particular, the `CGHcall R` program which includes a mixture model to infer CNV status [9]. Second, we build blocks/regions of consecutive clones with similar signatures. To perform this step the `CGHregions R` library was used [10]. Third, the association between the CNV status of blocks and the trait is assessed by incorporating the uncertainty probabilities in `CNVassoc` function. And fourth, corrections for multiple comparisons must be performed. We use the Benjamini-Hochberg(BH) correction [2]. This is a heuristic method that is robust against positive dependence and increasingly conservative as correlation increases.

To illustrate, we apply these steps to the breast cancer data studied by Neve et al. [6]. The data consists of CGH arrays of 1MB resolution and is available from Bioconductor `http://www.bioconductor.org/`. The authors chose the 50 samples that could be matched to the name tokens of caArrayDB data (June 9th 2007). In this example the association between strogen receptor positivity (dichotomous variable; 0: negative, 1: positive) and CNVs was tested. The original data set contained 2621 probes which were reduced to 459 blocks after the application of `CGHcall` and `CGHregions` functions as we illustrate bellow.

Table 1: Steps to assess association between CNVs and traits for aCGH

Step 1.	Use any aCGH calling procedure that provides posterior probabilities (uncertainty) (`CGHcall`)
Step 2.	Build blocks/regions of consecutive probes with similar signatures (`CGHregions`)
Step 3.	Use the signature that occurs most in a block to perform association(`multiCNVassoc`)
Step 4.	Correct for multiple testing considering dependency among signatures (`getPvalBH`)

The data is saved in an object called `NeveData`. This object is a list with two components. The first component corresponds to a dataframe containing 2621 rows and 54 columns with aCGH data (4 columns for the annotation and 50 log2ratio intensities). The second component is a vector with the phenotype analyzed (strogen receptor posistivity). The data can be loaded as usual

```
> data(NeveData)
> intensities <- NeveData$data
> pheno <- NeveData$pheno
```

The calling can be performed using `CGHcall` package by using the following instructions:

```
\dontrun{
#########################################################
### chunk number 1: Class of aCGH data
#########################################################
library(CGHcall)
Neve <- cghRaw(intensities)

#########################################################
### chunk number 2: Preprocessing
#########################################################
cghdata <- preprocess(Neve, maxmiss=30, nchrom=22)

#########################################################
### chunk number 3: Normalization
#########################################################
norm.cghdata <- normalize(cghdata, method="median", smoothOutliers=TRUE)

#########################################################
### chunk number 4: Segmentation
#########################################################
seg.cghdata <- segmentData(norm.cghdata, method="DNAcopy")

#########################################################
### chunk number 5: Calling
#########################################################
NeveCalled <- CGHcall(seg.cghdata)
}
```

This process takes about 20 minutes, but to avoid wasting your time, we have saved the final object of class `cghCall` that can be loaded as

```
> data(NeveCalled)
```

We can then obtain the posterior probabilities. `CGHcall` function does not estimates the underlying number of copies for each segment but assigns the underlying status: loss, normal or gain. For each segment and for each individual we obtain three posterior probabilities corresponding to each of these three statuses. This is done by executing

```
> probs <- getProbs(NeveCalled)
```

This is a dataframe that looks like this:

```
> probs[1:5, 1:7]
```

```
               Clone Chromo BPstart   BPend X600MPE X600MPE.1 X600MPE.2
RP11-82D16 RP11-82D16      1 2008651 2008651   0.022     0.932     0.046
RP11-62M23 RP11-62M23      1 3367844 3367844   0.022     0.932     0.046
RP11-111O5 RP11-111O5      1 4261844 4261844   0.022     0.932     0.046
RMC01P070   RMC01P070      1 5918606 5918606   0.022     0.932     0.046
RP11-51B4   RP11-51B4      1 6068980 6068980   0.022     0.932     0.046
```

This table can be read as following. The probability that the individual `X600MOE` is normal for the signature `RP11-82D16` is 0.932, while the probability of having a gain is 0.046 and 0.022 of having a loss.

In order to determine the regions that are recurrent or common among samples, we use the `CGHregions` function that takes an object of class `cghCall` (e.g. object `NeveCalled` in our case). This algorithm reduces the initial table to a smaller matrix that contains regions rather than individual probes. The regions consist of consequtive clones with similar signatures [10]. This can be done by executing

```
\dontrun{
library(CGHregions)
NeveRegions <- CGHregions(NeveCalled)
}
```

This process takes about 3 minutes. We have stored the result in the object `NeveRegions` that can be loaded as usual

```
> data(NeveRegions)
```

Now we have to get the posterior probabilities for each block/region. This can be done by typing

```
> probsRegions <- getProbsRegions(probs, NeveRegions, intensities)
```

Finally, the association analysis between each region and the strogen receptor positivity can be analyzed by using the `multiCNVassoc` function. This function repeatedly calls `CNVassoc` returning the p-value of association for each block/region

```
> pvals <- multiCNVassoc(probsRegions, formula = "pheno~CNV", model = "mult",
+     num.copies = 0:2, cnv.tol = 0.01)
```

Notice that the arguments of `multiCNVassoc` function are the same as those of `CNVassoc`. In this example, we have set the argument `num.copies` equal to 0, 1, and 2 that corresponds to `loss, normal, gain` status used in the `CGHcall` function.

Multiple comparisons can be addressed by using the Benjamini & Hochberg approach [2]. The function `getPvalBH` produces the corrected p-values

```
> pvalsBH <- getPvalBH(pvals)
> head(pvalsBH)
```

```
  region         pval     pval.BH
1    319 2.891862e-06 0.001324473
2    318 1.633799e-05 0.002494267
3    320 1.576279e-05 0.002494267
4    316 8.998845e-05 0.010303677
5      9 2.865773e-04 0.011217002
6    298 2.027325e-04 0.011217002
```

Table 6 in [3] can be obtained by typing

```
> cumsum(table(cut(pvalsBH[, 2], c(-Inf, 1e-05, 1e-04, 0.001, 0.01,
+     0.05))))
```

```
  (-Inf,1e-05] (1e-05,0.0001] (0.0001,0.001]   (0.001,0.01]    (0.01,0.05]
             1              4             27             64            117
```

4 Illumina data

In this section an example set of data from ILLUMINA will be analyzed. This data is included in the `CNVassoc` package, and is the same one as analyzed in the `CNVtools` package vignette [8]. The goal of this section will be to compare the results yielded by `CNVtools` in fitting the association model with those obtained with the `CNVassoc` function.

A first look at the data

```
> data(A112)
```

```
> head(A112)
```

```
         subject cohort        SNP0       SNP1       SNP2       SNP3
1 WTCCC01-11474A1    58C -0.12647400 -0.1214220 -0.1423570  0.0449446
2 WTCCC01-11474A2    58C -0.21574200  0.0265778 -0.0964269  0.0617480
3 WTCCC01-11474A3    58C -0.00150499  0.0820076 -0.2853430  0.1589580
4 WTCCC01-11474A4    58C -0.05538290 -0.1691450 -0.0592800  0.0264289
5 WTCCC01-11474A5    58C -0.12926900  0.2014540 -0.8474870 -0.2647420
6 WTCCC01-11474A6    58C -0.06209860  0.1826130  0.1245160 -0.1731720
       SNP4       SNP5       SNP6       SNP7       SNP8       SNP9
1  0.0259435  0.1351870  0.0746991  0.40581000 -0.18601600  0.0990579
2  0.1521360 -0.0445652 -0.3751110 -0.39122600  0.10114500  0.1816270
3  0.0320422  0.1823220  0.0699921  0.29014900  0.00885492 -0.0387201
4 -0.0208353 -0.2740840  0.0310302  0.20566300  0.12842100 -0.2219500
5 -0.0502723 -0.2150250 -0.2254730  0.00162372  0.08069250  0.0562238
6 -0.0870918 -0.0902743 -0.0634414 -0.80391700  0.37845800 -0.1880560
      SNP10      SNP11      SNP12       SNP13     SNP14      SNP15      SNP16
1 -0.1969750  0.0448241 -0.0193997  0.13117800 -0.163383  0.1545760  0.0253607
2  0.0688791 -0.1166620  0.0217019 -0.05719720 -0.138044 -0.0554405 -0.0536655
3  0.1131100  0.0609800  0.2402140  0.23635400 -0.111235  0.5082330  0.0272966
4 -0.2299260  0.0198905 -0.3210060  0.14955900 -0.534339 -0.7596830 -0.1940050
5 -0.0636589 -0.0433160 -0.5579070  0.13913000 -0.778225 -0.9224910  0.0343805
6 -0.1368910 -0.0779523  0.1212290  0.00857489  0.179257 -0.0675581 -0.1812210
      SNP17      SNP18     SNP19      SNP20      SNP21      SNP22      SNP23
1 -0.0560689 -0.0751385 -0.485160 -0.0288187 -0.1945410 -0.0456346  0.0929479
2  0.1212250  0.1018410 -0.200404 -0.1797650 -0.0456029  0.2835270 -0.0813351
3 -0.1532160 -0.1135340  0.183407 -0.0960403  0.1230410 -0.1076840 -0.0180287
4 -0.1035420  0.1661650 -0.318173 -0.7149550 -0.7436040 -0.2483910 -0.2552810
5 -0.0130905  0.1538550 -0.589194 -0.4773230 -0.6345150  0.1788480 -0.4428020
```

```
6 -0.1676740  0.3261350 -0.199970  0.0908316  0.1268390  0.1787620  0.1138070
      SNP24     SNP25      SNP26       SNP27      SNP28        SNP29      SNP30
1 -0.222375 -0.368043 -0.1448800 -0.00706918  0.0356588 -0.346104000 -0.1318280
2 -0.143908 -0.105819 -0.2330800 -0.07807670  0.0980952 -0.152811000  0.0728393
3  0.401502  0.240364 -0.1334340 -0.00942116 -0.0514102 -0.254315000  0.0708932
4 -0.106774  0.203908  0.3008000 -0.24017000  0.1681400  0.298436000  0.1303020
5 -0.124996 -0.191220 -0.1863940 -0.08408520 -0.2589270 -0.000203031 -0.0516899
6 -0.228779 -0.409863  0.0208064  0.01472170  0.2187790 -0.384239000 -0.0265277
       SNP31      SNP32
1  0.0140277  0.0583939
2 -0.2176910  0.0172098
3 -0.1251580 -0.1050300
4 -0.0139433  0.0432413
5  0.0381275 -0.0992932
6 -0.2410430 -0.0577618
```

In this case, intstead of having just one signal, a considerable number of them define a single gene. In CNVtools vignette [8] these are all summarized using principal components analysis, and the first component is taken in order to obtain one signal value per individual. The following steps to obtain peak intensities are the same as in [8].

To begin, load `CNVtools` package, since some function from it will be used to execute some previous steps in order to mimic the analysis performed in [8]:

```
> library(CNVtools)
```

4.1 Preparing signal data

The raw signal from all probes of the data is subtracted typing

```
> raw.signal <- as.matrix(A112[, -c(1, 2)])
> dimnames(raw.signal)[[1]] <- A112$subject
```

Then, the unidimensional data is summarized using principal component techinque from raw signal data

```
> pca.signal <- apply.pca(raw.signal)
```

In the article on `CNVtools` [1] it is suggested not to use this summarized intensity, `pca.signal`. Instead, the probability of occurrence of each of the 3 copy number status values (loss, normal and gain) is estimated after fitting a normal-mixture model to `pca.signal` using the function `CNVtest.binary` from CNVtools package.

```
> ncomp <- 3
> batches <- factor(A112$cohort)
> sample <- factor(A112$subject)
> fit.pca <- CNVtest.binary(signal = pca.signal, sample = sample,
+     batch = batches, ncomp = ncomp, n.H0 = 3, n.H1 = 0, model.var = "~ strata(cn)")
```

and after this, a linear discriminant analysis on raw signal data and these probabilities is performed

```
> pca.posterior <- as.matrix((fit.pca$posterior.H0)[, paste("P",
+     seq(1:ncomp), sep = "")])
> dimnames(pca.posterior)[[1]] <- (fit.pca$posterior.H0)$subject
> ldf.signal <- apply.ldf(raw.signal, pca.posterior)
```

4.2 Inferring copy number status considering batch effect

Once all signal probe intensities from the same gene have been summarized (`ldf.signal`), regardless of the techinque used, a normal mixture model is fitted using function `cnv` as already explained in Section 2. A possible batch effect in inferring copy number status has been considered, as mentioned

in [1]. Therefore, and in order to better mimic the example as presented in CNVtools vignette [8], copy number status is inferred taking into account the batches, simply by incorporating an argument to function `cnv`, called `batches`:

```
> CNV <- cnv(ldf.signal, batches = batches, num.class = 3, mix = "mclust")
> CNV

Inferred copy number variant by a quantitative signal
   Method: function Mclust {package: mclust}

-. Number of individuals: 2593
-. Copies 1, 2, 3
-. Estimated means:
      CNV 1   CNV 2  CNV 3
58C -1.9703 -0.2361 0.7752
NBS -2.1398 -0.1708 0.9074
-. Estimated variances:
     CNV 1  CNV 2  CNV 3
58C 0.0941 0.0941 0.0941
NBS 0.0847 0.0847 0.0847
-. Estimated proportions: 0.1524, 0.4973, 0.3503
```

In this case, the method "mclust" has been used in order to make the mixture model converge. Thus, a normal mixture is fitted separately per batch, and copy number status probability is updated pooling the copy number frequency among all batches. Notice that although specific means and variances are estimated per batch, only one pooled set of copy number frequencies is produced.

Also note that `plot` behaves slightly differently for CNV estimated taking into account the batch effect, drawing specific density curves and mean lines for each batch (see figure 7)

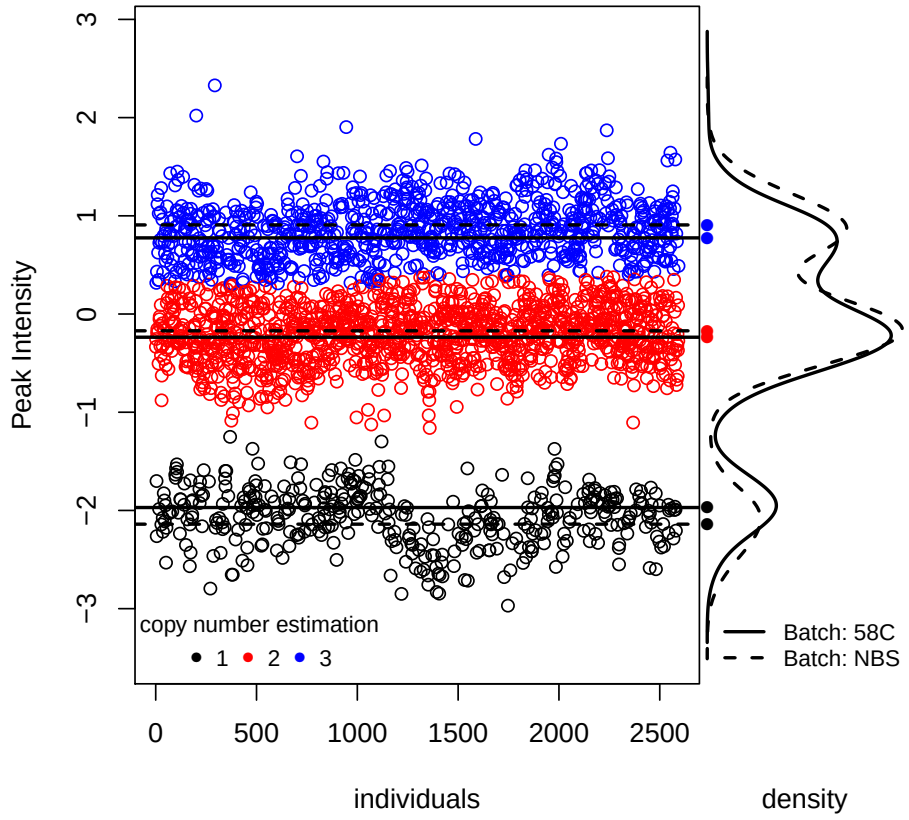


Figure 7: Signal distribution and inferred number of copies by batch

4.3 Association model: comparison with results from `CNVtools`

Now, the same batch variable will be the response as in [8], and an association model considering and additive effect test will be fitted. Since there are only 2 batches, a logistic regression will be performed. To compute the Likelihood Ratio Test on CNV:

```
> trait <- ifelse(A112$cohort == "58C", 0, 1)
> fit <- CNVassoc(trait ~ CNV, model = "add")
> CNVtest(fit, "LRT")

----CNV Likelihood Ratio Test----
Chi= 1.812608 (df= 1 ) , pvalue= 0.1781957
```

This results in a $\chi^2 = 1.81$ which does not differ greatly from the one given in CNVtools vignette [8] (1.55), neither being statistically significant.

And if a multiplicative model is assumed,

```
> fit <- CNVassoc(trait ~ CNV)
> CNVtest(fit, "LRT")

----CNV Likelihood Ratio Test----
Chi= 2.860054 (df= 2 ) , pvalue= 0.2393024
```

a χ^2 of 2.86 is obtained, similar to that in CNVtools-vignette [8] (3.11). Again, neither is statistically significant.

4.3.1 Power and computation time of CNVassoc and CNVtools

We simulated, under the same conditions as used by [1], a range of scenarios with different sample sizes, probe signal intensity distributions, etc., in order to explore the behavior of both methods when the copy number signals are not clearly separated. We observe that both methods performed well although CNVassoc outperforms CNVtools in the case of having a moderate number of individuals (e.g. 500) , see figures 8 and 9. However, we encounter and important problem of practical relevance related to convergence. CNVtools frequently fails to converge with moderate sample sizes: with 500 cases and 500 controls and $Q = 3$ [1], CNVtools failed to converge in more than 75% of the simulations and this failure rate reached 86% when $Q = 2.5$. Even with much larger sample sizes (2,000 cases and 2,000 controls) CNVtools failed to converge in 38% of the simulations when $Q = 2.5$. In constrast, CNVassoc converged in all scenarios with large sample size (2,000 and 2,000 controls) and with moderate sample sizes (500 cases and 500 controls) CNVassoc did not fail under low/moderate uncertainty $Q \geq 3.5$ and failed but much less than CNVtools when $Q \leq 3$, see table 2. Thus, for many studies being analyzed currently, CNVtools simply cannot provide a solution. When a solution is reached, the high rate of failure to converge raises questions about possible biases and imprecision of the results and, in any case, the solution is unlikely to be powerful enough to detect an association between copy number and phenotype.

We have also observed a marked difference in the speed of each procedure: when analyzing 10,000 CNVs in 2,000 cases and 2,000 controls, and with a $Q = 4$, CNVtools took 1 day and 17 hours to complete the analysis, whereas CNVassoc took just 90 minutes; with $Q = 3$, CNVtools took 6 days and 16 hours, but CNVassoc took only 2 hours.

[1] Q is the measure of uncertainty in inferring copy number status defined by CNVtools package (obtained by specifiyng the argument `type="CNVtools"` in `getQualityScore` function)

	$N = 2000$		$N = 500$	
Q	CNVassoc	CNVtools	CNVassoc	CNVtools
6.0	0	0	0	15
5.5	0	0	0	20
5.0	0	0	0	65
4.5	0	0	0	92
4.2	0	0	0	187
4.0	0	0	0	246
3.7	0	0	0	294
3.5	0	1	0	299
3.2	0	13	212	389
3.0	0	65	331	400

Table 2: Number of failed convergence simulations out of 500 using CNVassoc and CNVtools according to inferring copy number uncertainty Q and number of cases N.

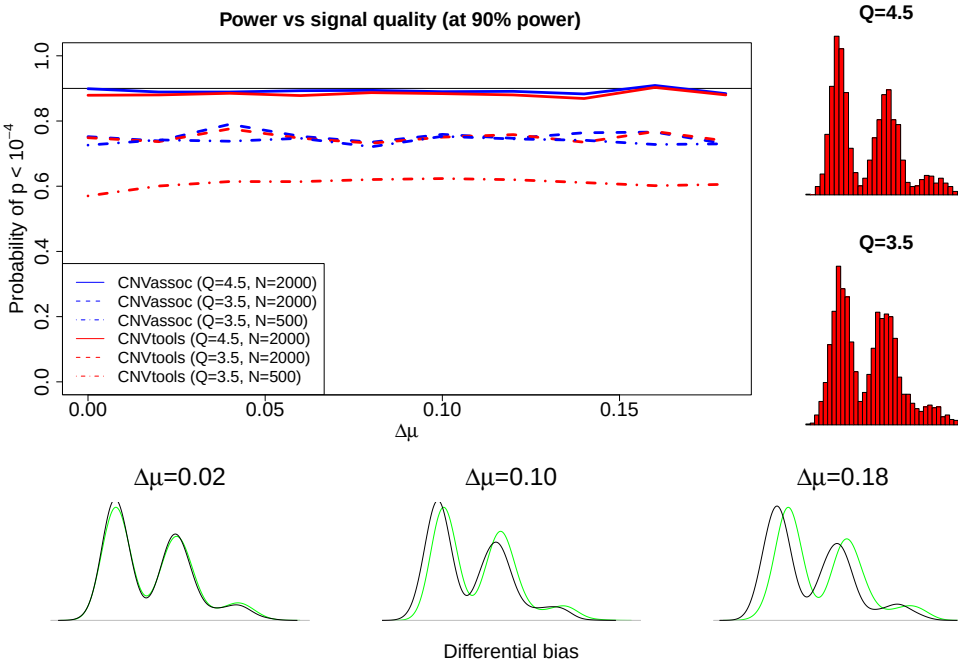


Figure 8: Power achieved by CNVassoc and CNVtools, depending on sample size, inferring copy number uncertainty (Q), degree of differencial bias ($\Delta\mu$) and sample size (N) under an scenario where the power to decect associated CNV is 90% if no inferring copy number uncertainty was present.

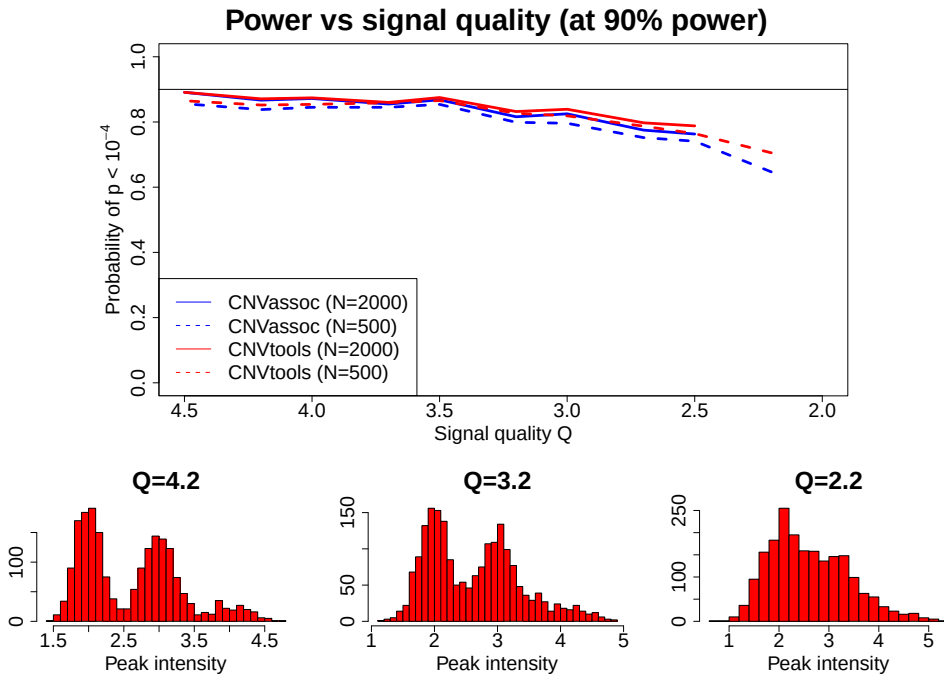


Figure 9: Power achieved by CNVassoc and CNVtools, with different values of copy number uncertainty (Q) and sample size (N) under an scenario where the power to decect associated CNV is 90% if no inferring copy number uncertainty was present.

5 Imputed data (SNPTEST format)

In this section we will show how `CNVassoc` can also be used to analyse SNP data when the SNPs have been imputed or genotyped with some degree of error. Notice that the same procedure can be applied to analyze data from Birdsuite/Canary software (developed by Broad Institute and available on http://www.broadinstitute.org/). An example from SNPTEST software (available on http://www.stats.ox.ac.uk/ marchini/software/gwas/snptest.html) has been incorporated in the `CNVassoc` package, but in the same format as used by IMPUTE software (downloable from SNPTEST website). IMPUTE is a program to infer a set non observed SNPs from other that have been genotyped, using linkage desequilibrium and other information, usually from the HapMap project (http://snp.cshl.org/). The data of the following example can be downloaded freely from the SNPTEST software website, and consists of a set of 500 cases and 500 controls, and 100 SNPs. For all of the SNPs the probabilities of each genotype is given, not the genotype itself, simulating having been obtained from IMPUTE. The names of the SNPs have been masked, as also the name of the disease.

Let's load the data. There are 2 data frames, one for cases and the other for controls

```
> data(SNPTEST)
> dim(cases)

[1]  100 1505

> dim(controls)

[1]  100 1505

> cases[1:10, 1:11]
```

```
   V1 V2    V3 V4 V5           V6           V7           V8          V9
1   1  1  1000  A  T 0.9959626125 0.0023620260 0.0016753615 0.992634932
2   2  2  2000  A  T 0.0765213302 0.0073893102 0.9160893596 0.027811741
3   3  3  3000  A  T 0.0050670931 0.0020722897 0.9928606172 0.009646064
4   4  4  4000  A  T 0.9920997158 0.0003108851 0.0075893991 0.012288000
5   5  5  5000  A  T 0.0048796013 0.0283927739 0.9667276249 0.990459821
6   6  6  6000  A  T 0.0029449045 0.9965970143 0.0004580812 0.993531065
7   7  7  7000  A  T 0.9844537961 0.0147126387 0.0008335652 0.003635098
8   8  8  8000  A  T 0.0002854996 0.0019421881 0.9977723123 0.005000345
9   9  9  9000  A  T 0.0052202003 0.0037747406 0.9910050592 0.003845385
10 10 10 10000  A  T 0.0145463505 0.9603995477 0.0250541018 0.010122825
           V10          V11
1  0.0003516265 7.013442e-03
2  0.0086429180 9.635453e-01
3  0.0026860830 9.876679e-01
4  0.9815783730 6.133627e-03
5  0.0092745162 2.656632e-04
6  0.0023760942 4.092840e-03
7  0.9945822710 1.782631e-03
8  0.0024962428 9.925034e-01
9  0.0011333510 9.950213e-01
10 0.9898094554 6.771937e-05
```

```
> controls[1:10, 1:11]
```

```
   V1 V2    V3 V4 V5           V6          V7          V8           V9
1   1  1  1000  A  T 9.822425e-01 0.003358295 0.014399242 0.9910275077
2   2  2  2000  A  T 1.333922e-02 0.969099360 0.017561421 0.0070884674
3   3  3  3000  A  T 3.989599e-03 0.004256366 0.991754036 0.0014208265
4   4  4  4000  A  T 3.406932e-03 0.007333515 0.989259553 0.0006075389
5   5  5  5000  A  T 9.881081e-01 0.010474830 0.001417104 0.9828012172
6   6  6  6000  A  T 3.595319e-03 0.990430376 0.005974305 0.0003284885
```

```
7   7  7  7000  A  T 6.072451e-05 0.997494894 0.002444382 0.0034642921
8   8  8  8000  A  T 6.322546e-03 0.006265613 0.987411841 0.0016109147
9   9  9  9000  A  T 3.073608e-04 0.007901964 0.991790675 0.0160832317
10 10 10 10000  A  T 9.748969e-03 0.978622828 0.011628203 0.0076508106
           V10         V11
1  0.001110983 0.007861509
2  0.028424366 0.964487167
3  0.984644304 0.013934870
4  0.997842168 0.001550293
5  0.011371321 0.005827462
6  0.995963534 0.003707978
7  0.989251393 0.007284314
8  0.006935266 0.991453820
9  0.981741626 0.002175142
10 0.973590298 0.018758891
```

The structure of the data is as follows:

- every row is a SNP
- the first 3 columns are the SNP identification codes,
- the 4th and 5th are the alleles.
- columns 6 through to the end provide the probabilities of each genotype, each group of 3 columns corresponds to one individual.

For example, the first individual in the data set of cases has probabilities of 0.996, 0.0024 and 0.0017 of having the genotypes for the first SNP of AA, AT and TT respectively. And the second individual has a probabilities of 0.0278, 0.0086 and 0.9635 of having the genotypes for the second SNP of AA, AT and TT respectively.

Of course, cases and controls must have the same number of rows, because the i-th row of cases and the i-th row of controls correspond to the same SNP.

First in order to use `CNVassoc` certain preliminary data management steps are needed. The goal is to have one matrix of probabilities with 3 columns corresponding to the 3 genotypes and 1000 individuals (500 cases plus 500 controls), for each of the 100 SNPs.

```
> nSNP <- nrow(cases)
> probs <- lapply(1:nSNP, function(i) {
+     snpi.cases <- matrix(as.double(cases[i, 6:ncol(cases)]),
+         ncol = 3, byrow = TRUE)
+     snpi.controls <- matrix(as.double(controls[i, 6:ncol(controls)]),
+         ncol = 3, byrow = TRUE)
+     return(rbind(snpi.cases, snpi.controls))
+ })
```

Now `probs` is a list of 100 components, each one containing the probability matrix of each SNP, and the first 500 rows of each matrix refers to the cases and the rest to the controls.

In this point, we can use `multiCNVassoc` as shown in section 3, to perform an association test of each SNP with case control status. But first, a casecontrol variable must be defined, which, in this example, will be a simple vector of 500 ones and 500 zeros.

```
> casecon <- rep(1:0, c(500, 500))
```

Now, we have the data ready to fit a model. For example, to compute the association p-value between every SNP and case control status assuming an additive effect:

```
> pvals <- multiCNVassoc(probs, formula = "casecon~CNV", model = "add",
+     num.copies = 0:2, cnv.tol = 0.001)
```

And, as in section 3, it is necessary to correct for multiple tests:

```
> pvalsBH <- getPvalBH(pvals)
> head(pvalsBH)

  region       pval   pval.BH
1      1 0.29083371 0.8400958
2      3 0.13235295 0.8400958
3      5 0.08296301 0.8400958
4      6 0.18826664 0.8400958
5      7 0.24967318 0.8400958
6      9 0.30321197 0.8400958
```

A frecuency tabulation of how many SNP achieve different levels of significance is obtained by:

```
> table(cut(pvalsBH[, 2], c(-Inf, 0.001, 0.01, 0.05, 0.1, Inf)))

(-Inf,0.001] (0.001,0.01]  (0.01,0.05]   (0.05,0.1]   (0.1, Inf]
           0            0            2            7           91
```

From these results, no SNP appears to be associated with case control status.

6 Other phenotype distributions

The examples of the previous section dealt with continuous normally distributed phenotypes, and binary traits. However, there are situations where we may be interested in associating CNV with a phenotype that is not normally distributed, or which is not a binary trait.

6.1 Poisson distributed phenotype

One example of a phenotype that doesn't fit with previous examples is a counting process, that could be the number of times that a patient replapses from a specific cancer. This could be modelled with a Poisson distribution.

`CNVassoc` incorporates the possibility to fit a Poisson distribution by specifying `family`="poisson". Also, `CNVassoc` has a function to simulate CNV data and Poisson phenotype. Therefore, in this section simulated data from this function will be analysed.

Data for 4000 individuals has been simulated under the following scenario:

- CNV copy number of 0, 1 and 2 with probabilities of 0.25, 0.5 and 0.25 respectively,
- CNV intensity signal means of 0, 1 and 2 for 0, 1 and 2 copies respectively,
- CNV intensity signal standard deviation of 0.4 for each copy,
- an additive effect with a risk ratio of 1.7 for each increment in copy number status,
- incidence of 0.12 of relapsing among individuals with zero copies (which means a probability of 0.6737 of having at least one relapse).

```
> set.seed(123456)
> rr <- 1.7
> incid0 <- 0.12
> lambda <- c(incid0, incid0 * rr, incid0 * rr^2)
> dsim <- simCNVdataPois(n = 4000, mu.surrog = 0:2, sd.surrog = rep(0.4,
+     3), w = c(0.25, 0.5, 0.25), lambda = lambda)
> head(dsim)

     resp cnv     surrog
446     0   1  0.1626554
2214    0   2  1.1287803
3535    1   3  1.4992945
3579    1   3  1.9024086
678     0   1 -0.2533025
2813    2   2  0.4879491
```

The result is a data frame with 3 variables, and as many rows as individuals. The description of these variables is:

- `resp`: response, distributed as a Poisson given the copy number status,
- `cnv`: the real copy number status, which, in practice, will be unknown and not considered in testing the association,
- `surrog`: the CNV intensity signal.

First an object of class `cnv` is obtained fitting a normal mixture to the intensity signal, as in section ... Note that to make the normal mixture converge "mclust" method is specified:

```
> CNV <- cnv(dsim$surrog, mix = "mclust")
> CNV

Inferred copy number variant by a quantitative signal
   Method: function Mclust {package: mclust}

-. Number of individuals: 4000
-. Copies 1, 2, 3
-. Estimated means: 0.0141, 0.9774, 1.9636
-. Estimated variances: 0.1631, 0.1631, 0.1631
-. Estimated proportions: 0.2479, 0.4804, 0.2717


-. Note: number of classes has been selected using the best BIC
```

Note that 3 copy number status have been inferred by BIC criteria. By default 1, 2 and 3 copies are assigned. To change the copy number status to 0, 1 and 2 copies, just change the `num.copies` attribute properly:

```
> attr(CNV, "num.copies") <- 0:2
> CNV
```

```
Inferred copy number variant by a quantitative signal
   Method: function Mclust {package: mclust}

-. Number of individuals: 4000
-. Copies 0, 1, 2
-. Estimated means: 0.0141, 0.9774, 1.9636
-. Estimated variances: 0.1631, 0.1631, 0.1631
-. Estimated proportions: 0.2479, 0.4804, 0.2717


-. Note: number of classes has been selected using the best BIC
```

Then, an association model with CNV and the phenotype assuming an additive effect is performed as usual, but specifying `family="poisson"` in the call to function `CNVassoc`:

```
> fit <- CNVassoc(resp ~ CNV, data = dsim, family = "poisson",
+     model = "add")
> coef(summary(fit))
```

```
            RR lower.lim upper.lim         SE     stat pvalue
trend 1.613005  1.450285  1.793982 0.05425561 8.811971      0
```

The same generic functions are appliable as for normal and binary traits. Note that, now, `summary` prints "RR" instead of "OR".

We can compare this to the "gold standard" model, where the phenotype is regressed to the true copy number status:

```
> fit.gold <- glm(resp ~ cnv, data = dsim, family = "poisson")
> table.gold <- c(exp(c(coef(fit.gold)[2], confint(fit.gold)[2,
+     ])), coef(summary(fit.gold))[2, 4])
> names(table.gold) <- c("RR", "lower", "upper", "p-value")
> table.gold
```

```
          RR        lower        upper      p-value
1.701183e+00 1.547603e+00 1.871468e+00 5.752637e-28
```

The confidence interval of the estimate contains the true relative risk, and the "gold standard" model gives similar results as the one fitted using `CNVassoc` function (latent class model).

Because the data has been simulated from a fixed scenario, we may be interested in comparing with an estimation made under a naive strategy, i.e. compared to fitting a standard log-linear Poisson model assigning the most probable copy number to each individual (best guess approach):

```
> fit.naive <- glm(resp ~ CNV, data = dsim, family = "poisson")
> table.naive <- c(exp(c(coef(fit.naive)[2], confint(fit.naive)[2,
+     ])), coef(summary(fit.naive))[2, 4])
> names(table.naive) <- c("RR", "lower", "upper", "p-value")
> table.naive
```

```
          RR        lower        upper      p-value
1.555179e+00 1.415058e+00 1.710412e+00 6.646768e-20
```

To sum up, table 3 gives the relative risk estimated under different models (gold standard, latent class and naive):

	RR	lower	upper
Gold	1.70	1.55	1.87
LC	1.61	1.45	1.79
Naive	1.56	1.42	1.71

Table 3: Comparison of RR estimated by the gold standard model, a latent class model (LC) and naive approach

6.2 Weibull distributed phenotype

Similarly to a Poisson distributed phenotype, we may be interested in fitting data that comes from a followed cohort, where we want to estimate associations of time to death or onset of a particular disease with copy number variant. Probably some individuals will be censored, i.e. at the end of follow-up they are alive or free of disease. As for classical survival analysis is important to take into account these censored individuals and not to remove them from the analysis.

Function `CNVassoc` can handle this situation, simply by specifying `family`="weibull" rather than poisson or gaussian. In considering censoring status, function `Surv` must be invoked in the left hand term of the formula argument (as for `coxph` function for example).

In this subsection we illustrate how to fit a model with time to event, possibly censored, by fitting simulated data, in a similar manner to the previous subsection (Poisson distributed phenotype), and using function `simCNVdataWeibull` implemented in the CNVassoc package.

The following scenario has been simulated for 5000 individuals:

- CNV copy number of 0, 1 and 2 with probabilities of 0.25, 0.5 and 0.25 respectively,
- CNV intensity signal means of 0, 1 and 2 for 0, 1 and 2 copies respectively,
- CNV intensity signal standard deviation of 0.4 for each copy,
- an additive effect with a hazard ratio of 1.5 for each increment of copy number status
- shape parameter of the weibull distribution equal to one,
- disease incidence equal to 0.05 (per person-year) among the population with zero copies.
- proportion of non-censored individuals (who suffered the disease during the study) of 10%.

```
> set.seed(123456)
> n <- 5000
> w <- c(0.25, 0.5, 0.25)
> mu.surrog <- 0:2
> sd.surrog <- rep(0.4, 3)
> hr <- 1.5
> incid0 <- 0.05
> lambda <- c(incid0, incid0 * hr, incid0 * hr^2)
> shape <- 1
> scale <- lambda^(-1/shape)
> perc.obs <- 0.1
> time.cens <- qweibull(perc.obs, mean(shape), mean(scale))
```

```
> dsim <- simCNVdataWeibull(n, mu.surrog, sd.surrog, w, lambda,
+     shape, time.cens)
> head(dsim)

         resp cens cnv     surrog
739  1.482852    0   1  0.1436988
1282 1.482852    0   2  0.8899417
1339 1.482852    0   2  1.6149953
872  1.482852    0   1 -0.2586166
3718 1.482852    0   2  1.2688898
123  1.482852    0   1 -0.9089759
```

The result is a data frame with 4 variables (one additional variable, compared to the Poisson example, that corresponds to censoring indicator), and, as before, as many rows as individuals:

- `resp`: time to disease (weibull distributed) or censoring (end of follow-up),
- `cens`: censoring indicator (0: without disease at the end of follow-up period, 1: with disease within the follow-up period),
- `cnv`: the real copy number status, which, in practice, will be unknown and not considered in testing the association,
- `surrog`: the CNV intensity signal.

As before, the CNV signal is fitted under a normal mixture distribution with function `cnv` and specifying the "mclust" method:

```
> CNV <- cnv(dsim$surrog, mix = "mclust")
> CNV

Inferred copy number variant by a quantitative signal
   Method: function Mclust {package: mclust}

-. Number of individuals: 5000
-. Copies 1, 2, 3
-. Estimated means: 0.0081, 0.9805, 1.9833
-. Estimated variances: 0.1663, 0.1663, 0.1663
-. Estimated proportions: 0.2439, 0.4947, 0.2615


-. Note: number of classes has been selected using the best BIC
```

As for the Poisson example, 1, 2 and 3 copy number have been assigned. So, we need to change the copy number status to 0, 1 and 2 copies, and we proceed as before:

```
> attr(CNV, "num.copies") <- 0:2
> CNV

Inferred copy number variant by a quantitative signal
   Method: function Mclust {package: mclust}

-. Number of individuals: 5000
```

```
-. Copies 0, 1, 2
-. Estimated means: 0.0081, 0.9805, 1.9833
-. Estimated variances: 0.1663, 0.1663, 0.1663
-. Estimated proportions: 0.2439, 0.4947, 0.2615


-. Note: number of classes has been selected using the best BIC
```

Then, an association model with CNV and the phenotype assuming an additive effect is performed as usual, this time specifying `family="weibull"`, and introducing the censored status using function `Surv` in the left hand side of the formula argument: `CNVassoc` function:

```
> fit <- CNVassoc(Surv(resp, cens) ~ CNV, data = dsim, family = "weibull",
+     model = "add")
> coef(summary(fit))

            HR lower.lim upper.lim         SE     stat       pvalue
trend 1.385556  1.205619  1.592348 0.07097498 4.594595 4.335896e-06
```

Again, the same generic functions are applicable as for normal, binary traits and poisson distributed phenotype. Note that, now, `summary` prints "HR" instead of "OR" (binary) or "RR" (poisson).

References

[1] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M. E. Hurles. A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40(10):1245–52, 2008.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57:289–300, 1995.

[3] J. R. Gonzalez, I. Subirana, G. Escaramis, S. Peraza, A. Caceres, X. Estivill, and L. Armengol. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformatics*, 10:172, 2009.

[4] J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nat Genet.*, 40(10):1253–60, 2008.

[5] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 41(3):334–341, 2009.

[6] R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N Bayani, J-P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W-L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, and J. W. Gray. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10:515 – 527, 2006.

[7] F. Picard, S. Robin, E. Lebarbier, and J. J. Daudin. A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3):758–766, 2007.

[8] V. Plagnol and C. Barnes. Cnvtools vignette. November 28, 2009.

[9] M. A. van de Wiel, K. I. Kim, S. J. Vosse, W. N. van Wieringen, S. M. Wilting, and B. Ylstra. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23(7):892–894, 2007.

[10] M. A. van de Wiel and W. N. van Wieringen. CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, 2:55–63, 2007.

6.3. Otras publicaciones relacionadas

En el siguiente artículo se ha usado el modelo propuesto en esta tesis doctoral (mediante las funciones del paquete `CNVassoc`) y en el que se descubrieron CNVs asociados a distintas enfermedades.

- *Influence of fetal glutathione S-transferase copy number variants on adverse reproductive outcomes.* (pág. 206).

DOI: 10.1111/j.1471-0528.2012.03400.x
www.bjog.org

Short communication

Influence of fetal glutathione *S*-transferase copy number variants on adverse reproductive outcomes

M Bustamante,[a,b] A Danileviciute,[c] A Espinosa,[a,d] JR Gonzalez,[a,d] I Subirana,[d,e] S Cordier,[f] C Chevrier,[f] L Chatzi,[g] R Grazuleviciene,[c] J Sunyer,[a,e] J Ibarluzea,[e,h], F Ballester,[i,j] CM Villanueva,[a,e] M Nieuwenhuijsen,[a,e] X Estivill,[b,e] M Kogevinas,[a,k]

[a] Centre for Research in Environmental Epidemiology, Barcelona, Spain [b] Genetic Causes of Disease Group, Genes and Disease Programme, Centre for Genomic Regulation, Barcelona, Spain [c] Vytautas Magnus University, Kaunas, Lithuania [d] Hospital del Mar Research Institute, Barcelona, Spain [e] Public Health and Epidemiology Network Biomedical Research Centre, Instituto de Salud Carlos III, Madrid, Spain [f] Inserm, IRSET, Université Rennes, Rennes, France [g] Department of Social Medicine, Faculty of Medicine, University of Crete, Heraklion, Greece [h] Subdirección de Salud Pública-Gipuzkoa, San Sebastián, Spain [i] Area de investigación en Ambiente y Salud, Centro Superior de Investigación en Salud Pública and Public Health and Epidemiology Network Biomedical Research Centre, Valencia, Spain [j] University of Valencia, Valencia, Spain [k] National School of Public Health, Athens, Greece

Correspondence: M Bustamante Pineda, Centre for Research in Environmental Epidemiology- CREAL, C. Doctor Aiguader 88; 08003 Barcelona; Spain. Email mbustamante@creal.cat

Accepted 21 April 2012. Published Online 7 June 2012.

A nested case–control association study was designed to investigate the influence of maternal and fetal copy number variants (CNVs) on reproductive outcomes. Genotypes of ten CNVs encompassing *GST* and *CYP* genes were assessed. Significant associations were only found for child CNV genotypes. In particular, the child *GSTM1* insertion allele was associated with prematurity protection (odds ratio, 95% CI: 0.67, 0.51–0.89; $P < 0.01$), whereas the child *GSTT2B* insertion allele was associated with an increased risk of being small for gestational age (odds ratio, 95% CI: 1.33, 1.07–1.67; $P = 0.01$). The study highlights the role of the fetal genome in prenatal development and also the need to analyse CNVs in a systematic manner.

Keywords Copy number variant, glutathione *S*-transferase, preterm, small for gestational age.

Please cite this paper as: Bustamante M, Danileviciute A, Espinosa A, Gonzalez JR, Subirana I, Cordier S, Chevrier C, Chatzi L, Grazuleviciene R, Sunyer J, Ibarluzea J, Ballester F, Villanueva CM, Nieuwenhuijsen M, Estivill X, Kogevinas M. Influence of fetal glutathione *S*-transferase copy number variants on adverse reproductive outcomes. BJOG 2012;119:1141–1146.

Introduction

A large proportion of perinatal and infant mortality is accounted for preterm and very-low-birthweight newborns. Moreover, these adverse reproductive outcomes lead to high risks of developing several disorders later in life.

The heritabilities for gestational age and birthweight have been estimated between 25% and 50%. Gestational age exhibits slightly lower heritability than birthweight, and, in contrast to birthweight, maternal genetic factors have been found to be more important in preterm delivery than fetal factors.[1] Several genes have been associated with birthweight-related outcomes and the only genome-wide association study performed until now revealed two new loci: *ADCY5* and *CCNL1*.[2] Association studies in candidate genes for preterm birth have been reviewed elsewhere.[3]

Given that both genetic and environmental factors are important in determining reproductive outcome susceptibility, genes involved in detoxification of xenobiotics have been extensively investigated. Glutathione *S*-transferases (GSTs) are phase II enzymes that catalyse detoxification of electrophilic compounds, endogenous or exogenous, by conjugation to glutathione.[4] In addition to the glutathione activity, some GSTs can act as peroxidades, isomerases and thiol transferases, and they can also participate in signalling processes. Although GST proteins exhibit some overlap, specificities have also been observed. While GSTs α, μ and π are close in terms of structure and substrate specificity, the θ family (GSTT) presents some particular

characteristics.[4] The GSTTs show preference for conjugation of gluthatione to small xenobiotics and they can bioactivate particular xenobiotics and produce metabolites with higher reactivity.[4]

Genetic polymorphisms in GST genes account for part of the variability in GST enzymatic activity observed between individuals. Some GST genes (*GSTM1*, *GSTT1* and *GSTT2B*) are known to be located in copy number variable (CNV) regions.[4,5] Effect on adverse reproductive outcomes of the *GSTT1* and *GSTM1* deletion polymorphisms has not been consistently replicated. The irreproducibility of the results might be attributed, in addition to epidemiological bias, to the noncomprehensive analysis of the CNVs. First, not all of the CNVs located in GST clusters have been analysed with respect to disease,[5] and second, in the association studies the exact number of copies has not been determined.

Here we undertook a comprehensive analysis of ten CNVs encompassing detoxification genes (*GST*s and *CYP*s [cytochrome P450]). Only four of them were common in the population and their effects on adverse reproductive outcomes were explored.

Methods

Study population

Study participants were part of four European birth cohorts that participated in the Health Impacts of Long-Term Exposure to Disinfection By-Products in Drinking Water (HIWATE) project: INfancia y Medio Ambiente, Spain (INMA), including three subcohorts: INMA-Sabadell, INMA-Gipuzkoa, INMA-Valencia; Rhea (Greece); Pelagie (Perturbateurs endocriniens: Étude Longitudinale sur les Anomalies de la Grossesse, l'Infertilité et l'Enfance, France); and Kaunas (Lithuania).

Samples and outcome definitions

Preterm birth was defined as being born before 37 completed weeks of gestation. Gestational age was calculated based on the last menstrual period date if the last menstrual period and ultrasound-based (<20 weeks) estimations were consistent by 7 days or fewer. If not, we used gestational age estimated by ultrasound. If neither of these measures was available, then we used the gestational age registered by the maternity records. Almost all the infants were late preterm and <10% of them were born before 32 completed weeks of gestation. Small-for-gestational-age (SGA) infants were defined as having a birthweight below the 10th centile based on existing local scales or based on new customised scales, depending on the availability of adequate growth curves. In the INMA and Kaunas cohorts, SGA was defined based on local standard scales, whereas customised scales were used in Rhea. A combination of local and customised scales was used in Pelagie. Controls, matched to cases by sex and country within each cohort, were randomly sampled. First, children that followed the inclusion criteria for controls (not preterm, not SGA, not large for gestational age) were preselected and grouped into subsets defined by country and sex. Then, from these subsets, a random sampling of children needed to equal the number of cases was performed using the sample command in STATA 8.0 (STATA Corp., College Station, TX, USA). The same controls were used for preterm and SGA analyses. Ethnicity was assessed with a self-administered questionnaire. Information on covariates was harmonised between cohorts and used for adjustment.

DNA extraction

DNA from cord blood or from maternal peripheral blood was extracted using different protocols (see Appendix S1), quantified using the PicoGreen dsDNA kit (Invitrogen, Carlsbad, CA, USA) and normalised to 40–60 ng/μl. A total of 2005 samples with available DNA were included in the study, 1003 maternal DNAs and 1002 child DNAs. Maternal DNA was not available in the Pelagie cohort.

Genotyping

Ten putative CNVs encompassing or near detoxification genes, *GST*s or *CYP*s, were identified using data from the Database of Genomic Variants, from the Wellcome Trust Sanger Institute database and from the literature. The CNVs were genotyped with the Multiplex Ligation-dependent Probe Amplification (MLPA) method (MRC-Holland, Amsterdam, the Netherlands) following the manufacturer's instructions with minimal modifications. MLPA data were normalised using the reference probes included in the assay and CNV status probability was inferred using the *mclust* algorithm implemented in the CNVAssoc package (R environment). To ensure the correct genotyping of the CNVs, we compared intertechnique and interlaboratory variability. Nine-hundred and ninety maternal DNAs (average call rate 98.7%) and 938 child DNAs (average call rate 93.6%) were successfully genotyped. Four samples out of 648 mother–child pairs (0.6%) showed Mendelian errors in *GSTT1* CNV (using a 0.9 CNV inference cut-off), so they were excluded from the analysis. Linkage disequilibrium parameters—the standardised coefficient of linkage disequilibrium (D') and the squared correlation coefficient measure of linkage disequilibrium between two loci (r^2)—were estimated using the Haplo.view programme. (See Supporting information, Appendix S1, for details on the MLPA genotyping process and validation.)

Statistical analysis

Families with genotypic data but lacking information of covariates were excluded (see Appendix S1). Hardy–

Weinberg equilibrium was explored in controls using a chi-square test. Logistic regression models were applied to the data under a codominant genetic model (and in some cases dominant and recessive models were fitted). Crude models were adjusted for cohort, whereas fully adjusted models contained, in addition to cohort, information about child sex, maternal age, maternal prepregnancy body mass index (BMI), maternal education, maternal smoking during first trimester of pregnancy and parity. As no large differences were observed between crude and adjusted models, only the adjusted models are shown. To take into account heterogeneity between cohorts (including genetic heterogeneity), random effects meta-analyses were performed. Statistical packages STATA 8.0, SPSS v17.0 and SNPassoc, CNVassoc and rmeta in R environment were used. A P value <0.05 was considered significant.

Results

Study samples

Complete data were available for 884 child DNAs and 913 maternal DNAs from four European birth cohorts. The main characteristics of the families are shown in Supporting information, Table S1. All covariates were similar in cases and controls, except for parity, maternal prepregnancy BMI and smoking during the first trimester of pregnancy. Preterm newborns were more frequent in second or subsequent pregnancies. In contrast, SGA children tended to be delivered to nulliparous mothers. Mothers who smoked tended to have smaller children. High maternal prepregnancy BMI was associated with preterm birth, whereas low BMI was associated with SGA.

CNV and haplotype frequencies

Seven CNVs in *GST* genes and three in *CYP* genes were evaluated. CNVs in the *CYP* genes (*CYP2D6*, *CYP2E1* and *CYP2A6*) were not further analysed because the genotyping classification quality score was not good enough. On the other hand, three of the *GST* CNVs (*GSTA2*, *GSTA3-A4* and *GSTM5*) were found to be monomorphic. Finally, four common CNVs (*GSTM4-5′*, *GSTM1*, *GSTT2B* and *GSTT1*) in two *GST* loci were analysed in relation to adverse reproductive outcomes. All four of these CNVs were in Hardy–Weinberg equilibrium in the controls, except for *GSTT1* which slightly deviated from the equilibrium in children (see Supporting information, Table S2). The *GSTM* CNVs were not in linkage disequilibrium (see Supporting information, Table S3, Figure S1a). In contrast, D' between *GSTT* CNVs was > 0.7 and r^2 was > 0.5 (Supporting information, Table S3, Figure S1b). Tag single nucleotide polymorphisms and *GSTT* CNVs showed an $r^2 > 0.6$.

CNV association analysis

Child *GSTM1* CNV was associated with preterm birth (Table 1). The *GSTM1* insertion allele conferred protection against preterm birth (odds ratio 0.67, 95% CI 0.51–0.89; $P < 0.01$; additive model). On the other hand, child *GSTT2B* CNV was associated with SGA and a trend was observed for preterm birth (Table 1). In particular, children bearing the GSTT2B insertion had an increased risk for being SGA (odds ratio 1.33, 95% CI 1.07–1.67; $P = 0.01$; additive model). Other CNVs were not associated with the outcomes. Random-effect meta-analysis for the estimates by cohort showed similar effects and P values for heterogeneity were 0.16 and 0.87 for child *GSTM1* and child *GSTT2B* CNVs, respectively (see Figure S2).

Regarding maternal genotypes, none of the CNVs was associated with the reproductive outcomes (see Supporting information, Table S4); however, maternal *GSTM1* CNV showed a trend similar to that described in children. When maternal genotypes were taken into consideration in the statistically significant models for child genotypes, no large differences in the estimation of the effects were detected (between 4% and 11%) (see Supporting information, Table S5).

CNV haplotype association analysis

We then tested the effect of child *GST* CNV haplotypes on adverse reproductive outcomes (see Supporting information, Table S6). An increased risk gradient was observed for each insertion allele in the *GSTT* locus. SGA cases were less frequent among those children bearing the deletion-deletion GSTT haplotype. A nominal association was found for preterm birth in the *GSTM* locus, but no clear pattern was observed.

Discussion

In the present study, we undertook a comprehensive analysis of CNVs in *GST* and *CYP* genes and, after filtering for quality control and CNV frequency, four common CNVs in or near *GST* genes were further explored in relation to adverse reproductive outcomes. We found that child *GSTM1* and *GSTT2B* CNVs, but not maternal CNVs, were associated with preterm birth and SGA, respectively. The haplotypic analysis suggested that not only the *GSTT2* CNV, but also the *GSTT1*, had an effect on SGA phenotype.

In general, mother or child *GSTM1* deletion has been reported to increase the risk for adverse reproductive outcomes, alone or in combination with certain exposures, and this is in the same direction as the association found in children in this study. According to these findings, the deletion of the *GSTM1* gene, and plausibly the decrease in detoxification capacity, conferred a higher risk for preterm delivery.

Table 1. Adjusted association analysis between child copy number variants and adverse reproductive outcomes*

Genetic model	Genotype	N** (control/case)	Preterm			N** (control/case)	Post-term		
			OR	95% CI	P value		OR	95% CI	P value
GSTM4-5′									
Codominant	0	147/72	1			147/134	1		
	1	186/87	0.93	0.62–1.39	0.72	186/141	0.8	0.57–1.11	0.18
	2	58/29	0.95	0.55–1.65	0.85	58/51	0.92	0.58–1.45	0.72
Codominant	0	170/105	1			170/160	1		
	1	168/68	0.64	0.43–0.94	0.02	168/126	0.83	0.60–1.15	0.27
	2	53/15	0.49	0.25–0.95	0.03	53/40	0.85	0.52–1.40	0.52
GSTM1									
Dominant	0	170/105	1			–	–	–	nt.****
	1 + 2	221/83	0.6	0.42–0.87	<0.01	–	–	–	
Additive	Linear trend	391/188	0.67	0.51–0.89	<0.01	–	–	–	nt.****
GSTT2B									
Codominant	0	143/53	1			143/91	1		
	1	190/109	1.49	0.98–2.25	0.06	190/171	1.37	0.97–1.94	0.08
	2	58/26	1.41	0.78–2.52	0.25	58/64	1.76	1.12–2.78	0.02
Dominant	0	–	–	–	nt****	143/91	1		
	1 + 2	–	–	–		247/233	1.46	1.05–2.03	0.02
Additive	Linear trend	–	–	–	nt****	391/326	1.33	1.07–1.67	0.01
GSTT1***									
Codominant	0	77/29	1			77/62	1		
	1	165/107	1.56	0.93–2.60	0.09	165/153	1.15	0.76–1.73	0.52
	2	149/52	0.84	0.48–1.46	0.54	149/111	0.95	0.62–1.46	0.81

*Adjusted for cohort, parity, maternal age, maternal prepregnancy BMI, child sex, maternal education, smoking during first trimester of pregnancy.
**Estimated number of individuals for each genotype.
***Not in Hardy–Weinberg equilibrium in the controls in child DNAs.
****Not tested because codominant *P* values > 0.1.

On the other hand, we found that child *GSTT2B* CNV insertion allele, and so the insertion of the *GSTT2B* gene, conferred a higher risk for SGA. The increased risk observed for the presence of *GSTT2B* gene could be explained by bioactivation processes.[4] Alternatively, the effect of the *GSTT2B* CNV might be indirect modifying the expression of nearby genes, as previously reported,[5] or it might tag other genetic variants in the region (see Figure S1b). In fact, the *GSTT2B* CNV insertion allele is correlated with the *GSTT1* CNV deletion allele,[5] and the *GSTT1* deletion allele has been reported to increase the risk of low birthweight in mothers, alone or in combination with prenatal exposures.

A strength of this study is that we tested both maternal and child genotypes, and with this, and the adjustment of one by the other, we tried to disentangle which genome is responsible for the phenotype.[1] Data from this study suggest that child *GST* CNV genotypes are responsible for the effects. However, a nonsignificant association for *GSTM1* CNV was detected in mothers. This can be just a consequence of the fact that mother and child share half of their genome. Detoxification during prenatal life depends on mother, placenta and embryo–fetal capacities. Although it is known that detoxification genes suffer changes in their expression and activity in mothers during pregnancy and at different stages of development in the fetus, not much is known about the expression of genes located within CNVs. In particular, Raijmakers et al.[6] found that the main *GST* gene expressed in embryo and fetus was *GSTP1*, followed by *GSTA* and *GSTM1*, and no expression of *GSTT1* was detected, but specimen donors were null for the *GSTT1* CNV. In addition to the detoxification activity, some *GST* family members have been involved in metabolism and signalling processes that might be relevant for fetal development.

Another important aspect of the study is that we have explored *GST* CNVs in relation to adverse reproductive outcomes in a comprehensive manner. We have analysed not only the largely explored *GST* CNVs, but also five other putative CNVs situated in *GSTT*, *GSTM* and *GSTA* loci. As far as we know, this is the first time that *GSTT2B* CNV has been studied in relation to disease.[5] Furthermore,

we have reported the CNV status as zero, one or two copies and this seems to be crucial given the linear trend found in the associations. We validated the CNV genotypes with different methods, genotype classification probabilities were considered using latent class statistical models, tag single nucleotide polymorphisms validated the genotyping, and a sensitivity analysis excluding low-quality genotypes was performed. Although quality controls were applied, we have to acknowledge that the deviation from Hardy–Weinberg equilibrium of *GSTT1* CNV in children, probably because of a lower classification score, could have had an impact on our results.

We designed a case–control study to increase the statistical power; however, the sample size was limited. Hence, we cannot discard the possibility that some associations have been overlooked, especially given the fact that some overlap exists between both phenotypes and that some nonsignificant associations were observed. Another limitation is the broad range definition used for preterm, which includes spontaneous preterm labour, delivery because of maternal or fetal infections and premature prelabour rupture of the membranes. We have not subdivided preterm infants according to clinical subphenotypes because of the limited sample size. Although different genetic factors might produce these subtypes of prematurity, the joint analysis allowed us to increase the statistical power to detect those genes that might play a general role in gestational age independently of the pathophysiological characteristics of each subtype. Finally data from different European cohorts have been pooled, and although no genetic test has been performed to deal with population substratification, women reporting not to be white European were excluded, controls were selected matched by country of origin within each cohort and all of the analyses were adjusted for cohort. Moreover, random-effect meta-analysis did not reveal any heterogeneity.

Conclusion

Child CNV genotypes in *GST* detoxification genes (*GSTM1* and *GSTT2B*), but not maternal genotypes, were associated with adverse reproductive outcomes. These data suggest a role for the fetal genome in prenatal development, and also highlights the need to analyse CNVs in a systematic manner improving genotype calling and exploring haplotypes.

Acknowledgements

The authors would like to acknowledge all the participants in the study. Part of the DNA extractions was performed at the Spanish National Genotyping Centre (CEGEN). The HIWATE consortium, which partially funded this project, consists of more participants than the current author list and we would like to thank them for their input.

Disclosure of interests

None of the authors have a conflict of interest.

Contribution to authorship

MB participated in the design of the study, prepared the DNAs, did the genotyping, performed the statistical analysis and wrote the first draft of the manuscript; AD prepared the DNAs, did the genotyping and participated in writing the manuscript; AE prepared the databases and assisted with the statistical analysis; JRG and IS assisted with the statistical analysis; SC, CC, LC, RG, JS, JI and FB provided data from the Pelagie, Rhea, Kaunas and INMA cohorts, and participated in writing the manuscript; CMV, MN and XE participated in the design of the study and in writing the manuscript; MK conceived the study, participated in the design and in writing the manuscript. All authors read and approved the final manuscript.

Details of ethics approval

Protocols of all studies were approved by local ethics committees. All the women in the study signed a consent form that included the use of genetic data. Standard procedures for the protection of confidential individual information have been applied. The study ethics complied with the Declaration of Helsinki. INMA-Sabadell: the research protocol was approved by the Hospital del Mar (IMAS) Bioethics Committee (20 July 2005). INMA-Valencia: the research protocol was approved by the Hospital Universitario 'La Fé' Bioethics Committee (29 October 2004, minutes N. 44). INMA-Guipuzcoa: the research protocol was approved by the Hospital Donostia Bioethics Committee (13 July 2005, minutes N. 7/05). Kaunas: the research protocol was approved by the Lithuanian Bioethics Committee (Protocol No 36224, 2006-07-10-32) and an oral informed consent was obtained from all women. Pelagie: Commission Nationale de l'Informatique et des Libertés (CNIL), 31 May 2002, Ref No.902076. Rhea: Ethical Committee, University Hospital of Heraklion, Crete, Greece, 26 February 2007, Ref No. 46/2007.

Funding

This work is partly funded by HIWATE (www.hiwate.eu). HIWATE is a three-and-a-half year Specific Targeted Research Project, funded under the EU Sixth Framework Programme for Research and Technological Development by the Research Directorate-Biotechnology, Agriculture and Food Research Unit (Contract no Food-CT-2006-036224). The INMA cohort is funded by grants from Instituto de Salud Carlos III (Red INMA G03/176, CB06/02/0041, FIS-FEDER 03/1615, 04/1509, 04/1112, 04/1931, 05/1079, 05/1052, 06/1213, 07/0314, 09/02647, FIS-PI041436, FIS-PI06/

0867, FIS-PI081151, FISS09-PS09/02311), Public Health and Epidemiology Network Biomedical Research Centre (CIBERESP) (AA08_012), Generalitat de Catalunya-CIRIT (1999SGR 00241), Departamento de Sanidad del Gobierno Vasco (BIOEF06/002), Diputación Foral de Gipuzkoa (DFG06/004), Conselleria de Sanitat Generalitat Valenciana and Fundació Roger Torné. The PELAGIE cohort is funded by Inserm, French Ministry of Health, French Ministry of Labour, ANSES, ANR and InVS.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Genomics region containing the CNVs analysed in *GST* genes: (a) *GSTM* and (b) *GSTT*.

Figure S2. Meta-analysis of child *GSTM1* and child *GSTT2B* CNVs in relation to preterm delivery and SGA, respectively.

Table S1. Main characteristics of the individuals included in the study by maternal and child DNA availability.

Table S2. Estimated genetic frequencies and Hardy–Weinberg equilibrium *P* value for children and mothers.

Table S3. Linkage disequilibrium parameters in the mothers and children.

Table S4. Adjusted association analysis between maternal CNVs and adverse reproductive outcomes.

Table S5. Adjusted associations between CNVs and reproductive outcomes in children after adjusting for maternal genotypes (mother–child paired samples).

Table S6. Adjusted association analysis between child CNV haplotypes and reproductive outcomes.

Table S7. Probes used in the MLPA genotyping assay.

Table S8. Number of copies for each CNV in HapMap samples obtained in this study (MLPA or PCR) and in others (Wellcome Trust Sanger Institute [iCGH] and Zhao et al. [PCR]).

Appendix S1. Methods.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author. ■

References

1 Lunde A, Melve KK, Gjessing HK, Skjaerven R, Irgens LM. Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent–offspring data. *Am J Epidemiol* 2007;165:734–41.

2 Freathy RM, Mook-Kanamori DO, Sovio U, Prokopenko I, Timpson NJ, Berry DJ, et al. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nat Genet* 2010;42:430–5.

3 Plunkett J, Muglia LJ. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Ann Med* 2008;40:167–95.

4 Josephy PD. Genetic variations in human glutathione transferase enzymes: significance for pharmacology and toxicology. *Hum Genomics Proteomics* 2010;2010:876940.

5 Zhao Y, Marotta M, Eichler EE, Eng C, Tanaka H. Linkage disequilibrium between two high-frequency deletion polymorphisms: implications for association studies involving the glutathione *S*-transferase (GST) genes. *PLoS Genet* 2009;5:e1000472.

6 Raijmakers MT, Steegers EA, Peters WH. Glutathione S-transferases and thiol concentrations in embryonic and early fetal tissues. *Hum Reprod* 2001;16:2445–50.